# Modeling the Process by Which People Try to Explain Complex Things to Others

**Gary Klein**, MacroCognition LLC, Wakefield, MA, USA, **Robert Hoffman**, Institute for Human Machine Cognition, Pensacola, FL, USA, **Shane Mueller**, Michigan Technological University, Houghton, MI, USA, and **Emily Newsome**, ShadowBox LLC, Dayton, OH, USA

The process of explaining something to another person is more than offering a statement. Explaining means taking the perspective and knowledge of the Learner into account and determining whether the Learner is satisfied. While the nature of explanation—conceived of as a set of statements—has been explored philosophically and empirically, the *process* of explaining, as an activity, has received less attention. We conducted an archival study, looking at 73 cases of explaining. We were particularly interested in cases in which the explanations focused on the workings of complex systems or technologies. The results generated two models: *local explaining* to address why a device (such an intelligent system) acted in a surprising way, and *global explaining* about how a device works. The examination of the processes of explaining as it occurs in natural settings revealed a number of mistaken beliefs about how explaining happens, and what constitutes an explanation that encourages learning.

**Keywords:** sensemaking, explaining, artificial intelligence

## INTRODUCTION

From time to time, we all need to help another person understand why something happened, why a machine behaved in an unexpected manner, or even how a complex device works. When we explain, we don't want to fashion an explanation that is too detailed, or one that lacks the necessary detail. How does that happen? How are we able to effectively explain something to another person? Consider an imaginary example of what might happen if you are driving and relying on a navigator in the passenger seat.

Navigator/Driver Dialog

The navigator is using a GPS aid embedded in a smart phone. As the driver, you think you need to continue going straight at the next intersection, but the navigator tells you, "Turn left here at the light." You might say, "Left?" and your tone of voice, in this one word, tells the navigator that you are surprised and perhaps skeptical. A good navigator will then try to explain to you the reason: the GPS is showing a red line ahead, but instead of giving these details, knowing the turn is coming up, the navigator simply says, "Traffic," perhaps pointing to the cell phone. And you are satisfied.

The navigator knows that you are surprised and knows that there is a need to rapidly explain so that you understand. In this dialog, you, as the driver, are the Learner and the navigator is the Explainer.

Currently, we can't have this type of dialog that is central to coordination, directly with our GPS devices, "Left?" "Traffic." But perhaps we can get closer to these exchanges if we appreciate how explaining happens.

This imaginary example was not part of the corpus of actual cases we examined, but we use this example at various points in this article because it is brief, clear, and easily remembered.

Our objective in the work reported here was to explore the nature of explaining, as distinct from generating explanations. Explanations are

Address correspondence to Gary Klein, MacroCognition LLC, 21 Crescent Street, Apt 1421, MA 01880, USA, gary@macrocognition.com.

*Journal of Cognitive Engineering and Decision Making*
Vol. 00, No. 0, Month XXXX, pp. 1-20
DOI:10.1177/15553434211045154
Article reuse guidelines: sagepub.com/journals-permissions

statements and have properties that have been explored philosophically and empirically. In contrast, explaining as an activity has received less attention.

The process of explaining occurs in various contexts including interpersonal dialog and interaction with technology. One of those contexts is the field of Artificial Intelligence (AI). The challenge of explaining how AI systems work has been a long-standing one, as seen in the pioneering works of William Swartout and others (Clancey, 1983, 1987, McKeown & Swartout, 1987; Swartout & Moore, 1993). They argued that an AI system needs to have within it an explainable model of the task, and also, a model of the user. The intelligent tutoring literature (e.g., Clancey, 1987; Forbus & Feltovich, 2001) describes a variety of efforts to capture the Learner's perspective and adapt the training accordingly.

In contrast to current Deep Net and Machine Learning systems, the early explainable systems had easy access to symbolic notation of knowledge, possibly making it easier to create human-meaningful accounts of the workings of the system. Yet many of those explanation systems failed to achieve what they had promised. Today's AI systems use calculational mechanisms that are much more opaque and may take substantial inferencing to be made sense of, and so the challenge is even greater.

To some extent, a similar argument could be made about how humans think about the reasoning of other humans. Nevertheless, Pearl (2018) notes that even though we have such a meager understanding of how our minds work, and how other people think, we can still communicate with each other, learn from each other, guide each other, and motivate each other. We can dialog in a language of cause and effect. In contrast, we cannot dialog with intelligent machines, and one reason is that they do not "speak" meaningfully about cause and effect. People communicate via a language of reasons, which is different from the AI language of variables and weights and correlations.

Looking across the broad and deep literatures, there are many divergent concepts and stances about what constitutes causation and causal reasoning. In philosophy, for example, some argue that an explanation is the output of a process that generates a "literal" description of reality whereas others see causal attributions and descriptions as a social construct (e.g., Collins, 1992). In the work reported here, we take no stance on the nature of causation. Rather, our focus is on the empirical description of causal reasoning.

Also, in philosophy there is a focus on the qualities of explanations that make them "good." This is taken largely from the perspective of philosophy of science, and thus includes such criteria as accuracy and completeness. We know from empirical research that such criteria do not map well onto human reasoning. For reviews, see Hoffman et al. (2011) and Hoffman et al. (2017).

In a review of the literature, Mueller et al. (2018) examined 743 articles, papers, and books having to do with explanation, covering a range of disciplines but focusing on reports of attempts to evaluate AI systems, including intelligent tutoring systems. This review found that while there was ample material on the process of generating explanations, the process of explaining has not been much studied. There are exceptions (e.g., Goguen et al., 1983), but our literature review found that explanations are often taken to be context-free and purpose-free statements, which can be evaluated in terms of factors such as clarity, comprehensiveness, and accuracy. In contrast, "explaining" is an interactive activity that involves the Explainer and at least one Learner. It is interactive in the sense that, to be effective, the process of explaining needs to take the Learner's perspective into account.

As part of the DARPA Explainable Artificial Intelligence (XAI) program, we conducted a thematic review of what happens when a person tries to explain the reasons for a decision or action in "real-world" settings, especially to explain the workings of a device to another person. Our goal for this study was to model how people engage in the process of explaining to others. Additionally, such a model might help the researchers seeking to enhance the explainability of AI systems. The study described in this paper relied on 73 textual accounts of explanation and did not include observations or interviews. We collected textual examples from books, magazines, newspapers, and social media, of cases in which people created written documents in an attempt to explain events and systems to readers. Our approach can be regarded as a way of expressing thematic analysis as described by qualitative researchers, for example, Grounded Theory (Corbin & Strauss, 1990).

The literature review performed by Mueller et al. (2018) found that most of the "theories" of explanation are narrowly focused on one of two contexts—teaching school children or philosophy of science (e.g., Toulmin's model of argumentation, 1958; Toulmin et al., 1984). Further, most things that would qualify as theories of explanation are simply taxonomies of properties, modes, or goals of explanation.

Existing psychological research on explanation typically does not invoke a process model, but asks questions such as "does explanation involve causal reasoning" or "do different modes or types of explanation work better" or "how does similarity of an explained cause impact the effectiveness of an explanation" (see extensive work by Lombrozo, 2006, 2010, 2011).

Most of this research is based solely in the laboratory, using contrived situations and nonexperts (e.g., school children). The goal of the present study is to understand explanation "in the wild."

## METHOD

### Corpus of Examples

We identified and examined 73 examples, some as complex as the Air France 447 disaster, others much simpler. We attempted to learn about the process of explaining from these examples. None of the examples came from direct observation of dialog in which one person attempted to explain the case to another person. Twenty-one of the examples came from Degani's (2004) descriptions of automation failures. Other cases came from news media or other published accounts (16), or from interviews we had conducted for other projects (9). A small number of explanations (2) were from the lead author's personal experience. Twenty-five of the examples came from the Reddit website "Explain Like I'm Five," which attempts to explain complex phenomena for unsophisticated audiences [https://www.reddit.com/r/explainlikeimfive/].

The cases included intelligent systems (e.g., IBM's Watson playing *Jeopardy*, AlphaGo playing Go), minimally intelligent systems (e.g., autopilots of commercial airlines and passenger ships, cruise controls for automobiles), mechanical systems (e.g., ceiling fans, motel alarm clocks, blood pressure monitors), and some decision-making events that did not involve machines (e.g., Magnus Carlsen's dramatic Queen sacrifice).

The process of creating this corpus of cases was opportunistic. In the absence of a model of the process of explaining, we wanted to cast a wide net. We had no basis for establishing criteria for acquisition and selection of cases and we did not want to narrow our corpus prematurely. We were not conducting a formal meta-analysis, in which we would need to set clear criteria for acquisition and selection in advance. Even a minimal criterion such as an account that was sufficiently detailed was too restrictive because we were also interested in poor explanations, which allowed us to examine how they were inadequate. This wide-net approach has served us well in previous naturalistic studies on the nature of decision-making (Klein et al., 2010) and insight (Klein & Jarosz, 2011). This wide-net approach is also consonant with qualitative research methodology. Unlike experimentation, where one has key questions in advance and seeks an answer, in qualitative research, one seeks good questions.

We were selective in that we wanted to include at least some cases involving intelligent systems, automated systems, and mechanical systems because this research was part of a DARPA program on Explainable AI. Other than that, the cases were all selected because they held promise for involving informative cases of explaining.

The researchers used the corpus of 73 cases to inform their understanding of broad features of explanations (e.g., the mechanisms frequently employed by Explainers, the presence or absence of mental simulation, and barriers/errors). Of the 73 total cases, 42 were "global" explanations, and 31 were "local." Local cases involve explaining why a specific outcome occurred, whereas global explaining is about general principles such as how a device works. The 73 cases are listed in the Appendix.

From the larger set, 26 examples stood out based on their richness and their judged potential

**TABLE 1:** Corpus of 26 Examples of Explaining Activities.

Global Explanations

- Why do Westerners and Arabs baffle each other in the way they think?
- Why do motel clock alarms sometimes fail to wake us up?
- Why do automobile cruise control systems sometimes run amok?
- Why do autopilots sometimes quit working with no warning?

Local Explanations

- Why did Watson give the answer "Toronto" in *Jeopardy*?
- Why are there maggots in my dead refrigerator?
- Why did Air France flight #447 crash?
- How did Magnus Carlsen come up with his dramatic Queen sacrifice to win a chess championship?
- Why did my GPS take me down an absurd route from San Francisco airport to Monterey?
- Why did a firefighter in LA County Fire Department claim that a newbie had an attitude problem?
- Why did the police officer shoot the innocent African-American shopper in Beavercreek Ohio?
- Why did the Department of Justice confrontation with David Koresh end in disaster?
- Why did CPT Rogers, of the USS Vincennes, shoot down an unarmed Iranian airliner in 1988?
- Why did the USAF F-15s shoot down two US Army helicopters over northern Iraq in 1994?
- Why did Korean Air Lines flight 007 get shot down?
- Why did the cruise ship *Royal Majesty* get grounded?
- Why did the automatic blood pressure machine fool the surgical team into thinking that a patient with low blood pressure actually had high blood pressure?
- Why did the airline pilot fail to arm the spoilers, resulting in a crash?
- Why did our ShadowBox instructions fail to get the effect we wanted in a military study of Good Strangers?
- How did the firefighter know to order his crew out of the burning building?
- Why did the British naval officer on board HMS Gloucester order the shoot-down of a new track during Desert Storm?
- Why did a missile battery bomb itself?
- Why was the German Blitzkrieg tactic so effective against the French in WWII?
- Why did the USN ship John McCain get into a collision?
- How did an engineer discover that a rogue train was causing disruptions in service?
- How did researchers discover the cause of Yellow Fever?

for helping us construct a model of explaining and we chose these for more careful study. These examples had more detail than the others. The detail included more causes and more interconnections between causes. The 26 selected cases are listed in Table 1, and we have framed them as questions. Questions are an obvious starting point for explanations—questions people don't know the answer to. Not everything needs explaining. The corpus included both local and global cases. Citations are provided in the Appendix.

### Analysis of the Case Materials

We studied the explanatory materials relying on our own inductive and abductive reasoning to detect themes across the incidents.

Because this was an exploratory project, a naturalistic investigation, we did not establish a set of analytical protocols in advance. We began with 20 coding dimensions, informed by our broad analysis of the larger corpus of 73 cases. During the course of coding the smaller set of

26 explanations, we discovered that most of the dimensions either overlapped or were ambiguous, and we reduced the original 20 categories to eight: the purpose of the explaining process, the trigger for the process, the Causal Palette (see below), the type of mental simulation, the number of entities and transitions involved in the mental simulation, the mechanism used for the explaining (analogs/comparisons, contrasts, diagrams, counterfactuals), and the use of tacit knowledge. See Table 2 for the coding categories, response options, and their corresponding definitions.

Two research assistants independently coded each of the cases. Inter-coder agreement was assessed, and modifications were made to coding criteria in order to resolve disagreements and improve clarity.

We calculated inter-coder agreement on the basis of how often coders were aligned in all dimensions across each individual coding dimension. Each of the eight dimensions included subcategories developed to express detail about the nature of each given explanation. Raters selected one or more of the descriptors in these subcategories from a set of three to nine options, or provided numerical ratings (e.g., 0–5) for each.

Overall inter-rater agreement was 71.2%. The rates of essential agreement are sufficiently high to instill reasonable confidence in the conclusions.

## RESULTS

The primary purpose of requesting an explanation was to correct or at least unearth flawed beliefs ($n$ = 25). We defined beliefs as causal connections between initiating conditions and outcomes. The trigger for local explanations was primarily surprise (92%). For example, why did Watson answer "Toronto," while playing Jeopardy? Why were there maggots in my dead refrigerator? Why did a GPS device send one of the authors down an absurd route? Why did a Beavercreek Ohio police office shoot an unarmed customer in a Wal-Mart store?

A lack of information triggered the explanation in nine cases, and in six cases, the explaining appeared to be pre-emptive, to prevent confusion. That is, the Explainer offered information in advance of a problem or confusion, anticipating that there might be a difficulty. Only three of the cases were deemed to be a fill-the-gap effort, which researchers defined as intending to add facts to what is already known.

Mental simulation, as expected, played a large role in the process of explaining; the Explainers, in building their stories, were seeking to help the Learners mentally simulate how the events led up to the surprising outcome. Causal chains (simple progressions from one state to the next) were found in 14 of the 26 cases, and networks of causes (descriptions involving multiple intersecting causes that do not line up in a chain) were found in 14 cases. We have observed (Klein & Crandall, 1995) that the mental simulation in good causal chain explanations involve at most three or four causes, and our data supported this hypothesis. We also hypothesized, based on the findings of Klein and Crandall, that mental simulations would have no more than six transitions, and we only found two cases with more than two transitions. The maximum number of causal chain transitions in our corpus of cases was three.

We also examined the type of reasoning used. We found that contrasts were the most frequent (found in 23 of the 26 cases). Diagrams were the second most common, appearing in nine of the cases. Table 3 shows a summary of all coding criteria for the 26 explanations.

We used these results in addition to our general analysis of the larger corpus of 73 examples to create conceptual models of the two types of explaining activities, local and global explaining.

Local explaining seeks to justify why specific actions were taken or decisions were made. In contrast, global explaining involves confusion or uncertainty, usually about how a system works—the explaining in these instances is not tethered to any particular incident. The local/global distinction formed a part of the initial rationale of the DARPA XAI program, and we found the distinction useful in formulating the models presented below. However, we also found that local explaining efforts generally invoked some global issues about how things worked. Therefore, the local/global distinction

**TABLE 2:** Coding Categories and Response Options.

| Coding Category | Response Options | Definitions |
|---|---|---|
| Purpose of explanation | Predicting | Allow the user to predict or forecast a future event or outcome. |
| | Trust | Gain the user's trust, persuade the user, or help user make a better judgment of when to trust the system. |
| | Correcting or unearthing flawed beliefs or assumptions | Identify and correct misconceptions. |
| | Evaluating a person | Evaluate a person's performance based on the sophistication of their explanation. |
| | Derivation/history | Provide historical background about how the present state came to be. |
| Trigger for explanation process | Surprise | Violation of expectancies. |
| | Ignorance | Curiosity. |
| | Pre-emptive | Anticipating future surprise or confusion. |
| | Fill-the-gap | Adding facts to something already known. |
| Causal Palette | Event/decision/forces | External act or event that impacted the situation. |
| | Missing data | Data that, if known, may have changed the course of the incident. |
| | Erroneous data | Data were wrong or not portraying reality. |
| | Flawed beliefs | A belief that was wrong or not applicable to the current situation. |
| | Mismatches to intelligent system | A mismatch in knowledge, goals, constraints, level of engagement, reasoning tactics, affordances, situation assessment, between intelligent systems (including humans). |
| Mental simulation type | None | |
| | Causal chain | The domino effect. |
| | Causal landscape | A tangle; causes that are interrelated. |
| Entities involved in mental simulation | Number of entities | Causes given in the explanation. |
| Transitions involved in mental simulation | Number of transitions between entities | Links or relationships between the causes. |
| Explanation mechanism | Analog/comparison | Analogy, metaphor, simile, or other comparison |
| | Contrast | X instead of Y. |
| | Diagram | Explanation portrayed in diagram form. |
| | Counterfactual | If X were true, Y could have happened. |

*(Continued)*

TABLE 2 (Continued)

| Coding Category | Response Options | Definitions |
|---|---|---|
| Tacit knowledge | Perceptual discriminations | Interpreting and differentiating between cues. |
| | Patterns | Trends that are seen across examples; features that commonly go together. |
| | Familiarity/abnormality | Understanding of what is "normal" and/or what deviations from normal look like. |
| | Mental models | How something works; how different parts are related to one another within a system. |
| | Mindsets | An overarching set of beliefs/attitudes. |

is not clear-cut. If the explaining process was centered on the incident, we classified it as local. If the explaining process centered on the details of how something worked, we classified it as global even if it invoked a specific incident.

## Local Explaining

People request local explanations when they want to know why something happened. Events did not go as expected, or a machine acted up for some reason. Hence, there is a surprise. By "surprise," we are referring to a trigger for sensemaking, as we have discussed previously (Klein et al., 2006).

Figure 1 presents a model describing the process of local explaining. The process is fairly straightforward—a surprising event engenders a need for something to be explained, leading to a diagnosis of this request, followed by a process of building and then packaging the explanation in the context of the Learner's background. We reviewed all 31 examples of local explaining from of our set of 73 cases to create a generic description of the process of explaining, as shown in Figure 1. The vast majority of the local explaining examples were triggered by a surprise.

Surprise is a function of the Learner's mental model. Some events may surprise one person and not another. A driver with no idea of the layout of a city may not blink when the navigator says to turn left at the next light. In contrast, a driver who knows the neighborhood will be surprised and is likely to want the directive explained. The Explainer will need to take the Learner's perspective into account. Is the driver saying "Left?" because it was noisy and he/she wasn't sure of what the navigator said? Or is the driver questioning the directive? Here, the explanation is the dialog between the Learner and the Explainer, and not simply a set of statements.

In many of the examples we studied, the Explainer began by trying to diagnose the reason for the inquiry. Since an assumption was violated, the Explainer tried to determine which assumption was wrong, in order to correct it. This process is very focused and can be very brief, as opposed to presenting a lengthy account of a system or a situation and then trying to extract the gist.

Figure 1 shows the Learner's status at the left and lists some of the different features the Explainer might want to take into account about the Learner. If you are explaining something to the Learner, you may want to consider the richness of the Learner's mental model and simplify your account accordingly. And along with the Learner's mental model, you may want to determine the Learner's mindset: the beliefs in the mental model that are framing the way the Learner approaches the task. You may want to consider the Learner's goals in requesting the explanation in order to highlight issues that will be most relevant to the Learner. You might consider the time pressure that your conversation is under in order to determine whether to truncate your comments. You may try to anticipate any possible common ground confusions so that you can take extra care in defining the

**TABLE 3:** Frequency of All Coded Categories.

| Coding Category | Number of Explanations | Percentage of Total |
|---|---|---|
| *Purpose of explanation* | | |
| Predicting | 7 | 26.9% |
| Trust | 14 | 53.8% |
| Correcting/ unearthing flawed beliefs | 23 | 88.5% |
| Evaluating a person | 1 | 3.8% |
| Derivation/history | 17 | 65.4% |
| *Trigger* | | |
| Surprise | 22 | 84.6% |
| Ignorance | 9 | 34.6% |
| Pre-emptive | 6 | 23.1% |
| Fill the gap | 3 | 11.5% |
| *Causal Palette* | | |
| Events/decision/ forces | 20 | 76.9% |
| Missing data | 15 | 57.7% |
| Erroneous data | 7 | 26.9% |
| Flawed beliefs | 20 | 76.9% |
| Mismatches to intelligent system | 23 | 88.5% |
| *Mental simulation type* | | |
| None | 7 | 26.9% |
| Causal chain | 14 | 53.8% |
| Causal landscape | 14 | 53.8% |
| *Entities involved in mental simulation* | *2.42 | |
| *Transitions involved in mental simulation* | *0.50 | |
| *Explanation Mechanism* | | |
| Analog/ comparison | 5 | 19.2% |
| Contrast | 23 | 88.5% |
| Diagram | 9 | 34.6% |
| Counterfactual | 5 | 19.2% |
| *Tacit Knowledge* | | |
| Perceptual discriminations | 5 | 19.2% |
| Patterns | 10 | 38.5% |
| Familiarity/ abnormality | 15 | 57.7% |
| Mental models | 23 | 88.5% |
| Mindsets | 14 | 53.8% |

*Note*: The values in this table reflect codes generated by one or both raters.
*These values are averages across all explanations. Codes for each category were not mutually exclusive.

terms you are using. The perception of common ground will shape the amount of detail provided during the explaining process (see Klein et al., 2005, for a discussion of common ground and coordination). You may try to determine how the Learner is assessing the situation and also the Learner's role in the activity, so that you can provide background information as necessary. Thus, the process of explaining to another person is more complicated than merely issuing an explanatory statement.

The components shown in Figure 1 are features and processes, and the links between them display the flow of influence to enable the process of explaining.

The process of explaining is even more difficult when writing to an audience instead of having a face-to-face dialog with another person. Our cases came from published materials and the authors had to make inferences about the readers, for example, their knowledgeability, goals, and so forth. However, the model shown in Figure 1 is intended for interpersonal dialog. For face-to-face conversations, it may seem daunting for the Explainer to take into account the Learner features listed in Table 1, and we think it unlikely that Explainers are consciously and deliberately checking for each of these features—we expect that this kind of perspective taking is accomplished intuitively and that there are individual differences in sensitivity to Learner features.

*Learner status.* The Explainer appeared to take the Learner's features into account in 26 of the 31 instances of local explaining. In all of the 26 instances in which Learner features were

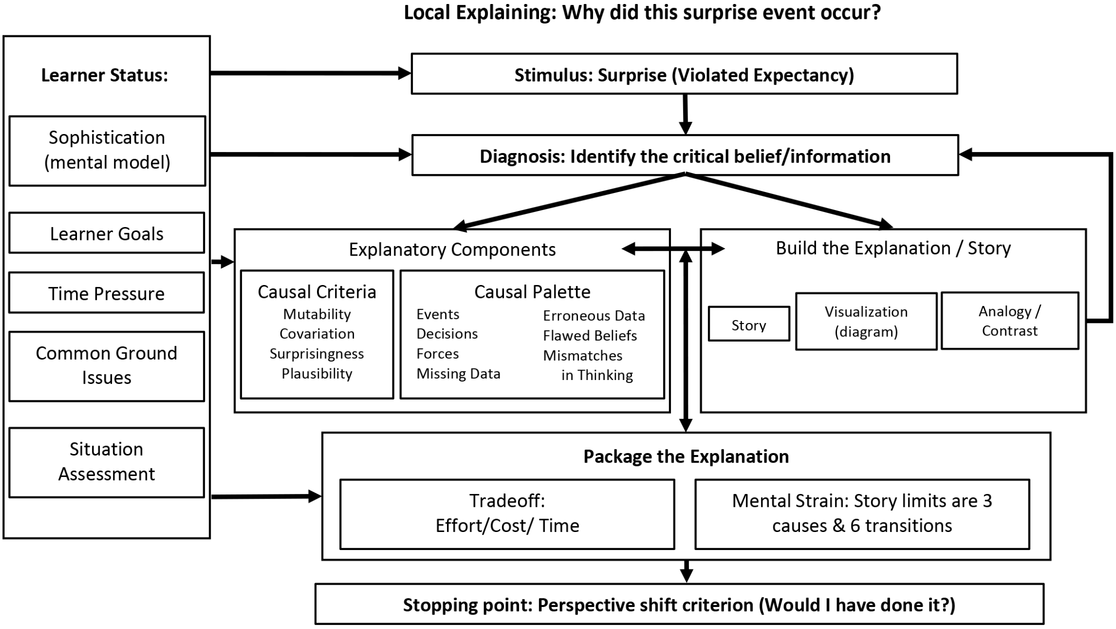
Local Explaining: Why did this surprise event occur?
Learner Status:
Sophistication (mental model)
Learner Goals
Time Pressure
Common Ground Issues
Situation Assessment
Stimulus: Surprise (Violated Expectancy)
Diagnosis: Identify the critical belief/information
Explanatory Components
Causal Criteria
Mutability
Covariation
Surprisingness
Plausibility
Causal Palette
Events
Decisions
Forces
Missing Data
Erroneous Data
Flawed Beliefs
Mismatches in Thinking
Build the Explanation / Story
Story
Visualization (diagram)
Analogy / Contrast
Package the Explanation
Tradeoff: Effort/Cost/ Time
Mental Strain: Story limits are 3 causes & 6 transitions
Stopping point: Perspective shift criterion (Would I have done it?)

*Figure 1.* Model of the process of local explaining.

considered, it seems to have been the richness of the Learner's mental model that was taken into account, and not the other factors shown in Figure 1. For example, in explaining why someone found maggots in a refrigerator that stopped working while the owner was on vacation, the Explainer addressed the assumption that the refrigerator was airtight. Even though the Learner did not state this assumption in the original question, the Explainer was able to infer the Learner's mental model of how a refrigerator works. All of the cases came from published accounts rather than observations of interactions, so factors such as time pressure and Learner goals and situation assessment and common ground did not come into play. We have included these factors in Figure 1 because we hypothesize that they may be very relevant during actual dialogs, as between two people or between an intelligent system and a user.

*Diagnosis.* After the request is triggered, the Explainer next has to diagnose what is behind it. This step is critical to the process of local explaining. The trigger is a violated expectation, and the Explainer tries to identify what was that expectation and why was it violated. In so doing, the Explainer simplifies his/her task. All the Explainer needs to do is provide information about how the Learner's expectation or assumption was wrong, or was limited, and what is a more accurate belief. The Explainer does not have to give a full account of how the device works or why the event unfolded as it did.

Given that our cases came from published materials, there was no face-to-face diagnosis. The diagnosis required the author to infer confusion. For example, Gladwell's account of David Koresh assumed that most of his readers would be as bewildered as he had been by Koresh's statements and actions, and he diagnosed the readers' (and his own) confusion as stemming from ignorance about the Branch Davidian movement.

Once the Explainer diagnoses the reason for the mismatch, the Explainer can use this reason as a focused explanation. A focused explanation is a much easier task than trying to present a comprehensive overview of all the causes and components involved. The driver says "Left?" and the navigator (the Explainer) gives a one-word response: "Traffic."

All of the 31 cases of local explaining appeared to depend on some level of diagnosis, and none of them tried for unnecessarily detailed accounts. However, in only three of the cases was there the “nutshell” account such as the driver/navigator interaction, “Left?” “Traffic.” We believe that the reason that these nutshells did not occur more often is that the cases we examined were too complex—they involved several causes, sometimes portrayed as a story and other times as a network. Nutshell responses would only occur when there was a single cause for a surprising event, or when the Explainer has some other reason to explain simply and rapidly, as opposed to complex incidents such as the Air France 447, example which is a blend of a story and a network, therefore not a straightforward causal chain.

The Air France 447 tragedy resulted in 228 deaths. The airplane took off from Rio de Janeiro, Brazil on June 1, 2009, on its way to Paris, France. However, 3 hr later, it crashed into the Atlantic Ocean. Some wreckage and two bodies were recovered in a few days, but the black boxes, the remaining bodies, and the bulk of the wreckage were not located for another 2 years. Why did the airplane crash? The explanation that the aircraft stalled begs the question of why it stalled. Many aspects of the pilot decision making are unclear and ambiguous, as described in the final report released in 2012 by BEA, France’s Bureau of Enquiry and Analysis for Civil Aviation Safety, and it is unlikely that a definitive explanation will ever be achieved. One of the prime causes was that the pitot tubes used to measure airspeed had iced up, and when airspeed information was lost, the autopilot system turned off and other instruments became unreliable, leading to confusion of the aircrew and a fatal decision to climb, resulting in the stall. What remains unknown, and subject to debate, is why the crew decided to climb and why the pilot flying seemed oblivious to the indications that the airplane was stalling.

In the absence of a conclusive account of the accident, a number of different explanations have been advanced for what went wrong. For example, Palmer (2013) focused on poorly trained pilots, and there is no denying that the pilots failed to understand the situation and as a result made poor decisions.

For the purpose of illustrating the model of local explaining shown in Figure 1, we examined a different account, one that suggests that the pilot error was itself caused by misunderstanding the intelligent technology that supposedly made the airplanes stall-proof (Sarter, personal communication, June 11, 2021). Why would the pilot climb so steeply without sufficient concern about stalling the airplane? One possibility is that the pilot flying had an erroneous belief that it was impossible to put this airplane into a stall. The manufacturers of the Airbus A330 had made this assurance. However, the mechanisms that were designed to prevent stall were valid only as long as the airplane sensors were working correctly. In this incident, when the pitot tubes had iced up and the airplane lost its ability to gage airspeed, it did become vulnerable to stalling but the pilot flying the airplane may have been unaware that it could now go into a stall. This account of the incident illustrates how a surprising event gave rise to explanatory effort. It illustrates how the diagnosis centered on a mistaken belief (that the airplane could not stall) and identified the condition that allowed a stall (the icing that affected the critical airspeed sensors.) A more detailed attempt to explain the accident identified seven distinct causes that resulted in confusions that chained and intersected. An event that first seemed to make no sense came into focus.

Further, the Air France 447 incident was a blend of story and network of causes. The basic causal chain at first seems straightforward but as the explaining process adds more details and causes, it turns into a network of causal factors.

Several of the examples presented below will further illustrate the Diagnosis step in Figure 1: The Wal-Mart shooting, the grounding of the Royal Majesty, the understanding of how maggots got into a freezer compartment. As we describe these examples, we will describe the Diagnostic step in each.

*Explanatory components*. Next comes the Explanatory Components, consisting of two sub-components that contribute to the content of how the local event is explained: the Causal Criteria for what can count as a cause and a Causal Palette consisting of types of causes that are often invoked in explaining something. In explaining something we offer causes. The Explanatory Components shown in Figure 1 refer to what counts as a cause and what kinds of causes might be invoked.

Hoffman et al. (2011) reviewed the literature on criteria for causes and identified four criteria. We are not claiming that people systematically search for causes by using these features—we are presenting the features so that readers will appreciate why certain issues are considered as possible causes and others aren't.

The four features are : mutability (what in this incident can be reversed to prevent the outcome?), covariation (what has varied in line with the effect?), surprisingness (e.g., inconsistency, connections, false assumptions), and plausibility, a feature that does not come into play in nominating causes but does come into play to assess causes and explanations—the potential cause has to seem plausible for producing the surprising outcome.

The Explanatory Components part of Figure 1 also includes a "Causal Palette." Reviewing the 31 incidents that involved local explaining, we identified several factors that are often cited as plausible causes for an event. We can consider these as a Causal Palette.

Just as artists will mix a custom set of colors on their palettes that get re-used for a particular painting, the Causal Palette is a set of the types of causes that get re-used for explaining in a particular situation. We do not claim that the Causal Palette shown in Figure 1 is complete. We have simply compiled a set of causes that are often cited. When an Explainer searches for reasons why a Learner has become confused, there is a good chance that the Explainer will consider topics from this Causal Palette. The Explainer is not sorting through the Causal Palette in the hopes of finding good candidates. Rather, the Causal Palette is the Explainer's attempt to describe which types of events are usually nominated as possible causes.

The items in the Causal Palette include: Events, Decisions, Forces, Missing data, Erroneous data, Flawed beliefs including mode errors, and Mismatches in thinking with that of an intelligent system (including another person). For example, in the GPS example described in the introduction, the causes include a flawed belief (the driver assumed they would be continuing straight), and an event (the traffic shown on the GPS device). Or consider the Air France 447 case. The surprise was that the PF (pilot flying) decided to climb steeply, resulting in a stall in an airplane that was never supposed to stall and a resulting crash.

To explain this event, we diagnose the causes, not all the causes of the event (which would generate a very large and complex influence diagram) but the causes that also created the surprise. One cause was an event: the frozen pitot tubes. Another cause was the erroneous and missing data, given that the airplane was no longer capable of assessing its airspeed. Another cause may have been a flawed belief of the pilot that the aircraft could not stall. These illustrate the general causal features invoked in accidents of this nature.

Our corpus included several examples involving erroneous data. In the case of the Wal-Mart shooting in Ohio, a police officer burst into a Wal-Mart and fatally shot a Black customer without even bothering to ask the customer any questions. How could this happen? The subsequent diagnosis explained that the police officer mistakenly believed the dispatcher had warned him of an active shooter, so the officer was following a protocol for handling active shooter. In another case, the cruise ship "Royal Majesty" grounded itself in clear weather. How could that occur? The diagnosis was that the cable attaching the captain's instruments to the GPS had come loose, and he was getting an erroneous picture while believing that it was accurate. In the case of the discovery of how Yellow Fever is transmitted, the "conclusive" experiment had been done and ruled out mosquitoes as a possible cause of transmission, but the experiment was flawed because it didn't take into account the 12 -day latency period following a mosquito bite.

The Air France flight #447 crash, described earlier, fits the category of flawed beliefs. The pilot flying the airplane may have believed the plane could not stall, but when the pitot tubes froze, the autopilot kicked off and the plane actually became vulnerable.

Regarding the last item in the Causal Palette, Mismatches, the first case shown in Table 1 refers to the IBM program Watson, which won a Jeopardy contest. Watson flubbed its final Jeopardy response to a clue "U.S. Cities": "Its largest airport is named for a World War II hero; its second largest, for a World War II battle." Watson responded, "What is Toronto?" which makes no sense at all. An Explainer might include this incident to illustrate that Watson obviously does not "think" the same way humans think—this is the mismatch that the story exemplifies.

In addition, we found mismatches in perspective. How do we diagnose the reason why another person or an intelligent system made a surprising decision or made an unexpected recommendation? When another person or intelligent system acts in a way that surprises us, the reason may be that that person/system has a different perspective than we do. We refer to these issues as a Perspective Mismatch (PM). The Explainer will have to diagnose one or more PM issue and determine if the Learner is mistaken or if, perhaps, the other person or the system is acting incorrectly. Then the Explainer will explain the reason for the mismatch to the Learner. We have identified seven PM issues—reasons that can explain the mismatch that resulted in a surprise: asymmetrical knowledge (we may know something the other entity does not, or vice versa), goals (we may be pursuing different goals), constraints, level of engagement, reasoning tactics, affordances (we may be aware of affordances that the other entity hasn't considered, and vice versa), and situational understanding.

*Building the story*. In parallel with identifying the causes to go into a story is the process of building the story around the causes—the way the explanation is communicated. The simplest version is for the Explainer to identify and name a single cause. In the GPS case, the navigator (the Explainer), just said "Traffic."

Most of the time, the Explainer will need to formulate a more elaborate account in explaining an event. Frequently, this account will take the form of a story. A typical story takes the form of a chain, one cause/event leads to a second, and then to a third, and to an outcome.

In building the story, plausibility comes into play. Each state transition in the chain has to plausibly follow from the previous state. The Learner needs to imagine how he/she would make the transitions. If plausibility is breaking down, the explanation is seen as problematic.

For example, consider this incident from our corpus. A couple went on a 2 -week vacation and when they returned home, to their surprise (and disgust) they discovered maggots and dead flies in the freezer compartment of their sealed refrigerator. How could that have happened? The diagnosis centered on the fact that in their absence the refrigerator had stopped working. As a result, the food in the refrigerator spoiled, which smells terrible to us but smells like perfume to a fly. Now it is starting to make sense. The flies must have laid eggs that hatched into maggots—the larvae can hatch into maggots in 24 hr. But there is a second mystery: how had flies gotten into the refrigerator? What do we know about a refrigerator that could attract and permit flies?

In this case, three different stories were developed. One story was that flies may have found a tiny crevice where the refrigerator door wasn't perfectly sealed. A second story was that the flies got in through the ice dispenser—that's another entry point because the flap-gate is weaker than the door seal. And the largest number of flies/maggots were found in the vicinity of the icemaker. A third story was that the flies never found an entrance into the refrigerator—this story centered around a picnic that the couple attended before they left town. The flies might have deposited eggs on some of the picnic food that was left unguarded, and the food was placed in the freezer and then the eggs hatched when the refrigerator stopped working. Thus, we have three stories, each composed of simple chains of events: a break in the door seal, a weak flap by the icemaker, and carelessness during a picnic.

With more complex cases, the Explainer may shift from story building to a Causal Landscape (Klein, 2018) or some other visual representation of a larger number of causes that operate in parallel and also intersect. Sometimes, the Explainer will draw a diagram to show the new belief/assumption. Explaining can take other forms, such as using a contrast to illustrate how the current situation is different from one that seems similar, or offering an analog to get the point across.

*Packaging the explanation*. Next, the Explainer will give some thought to packaging the explanation—executing the explanation. The context, along with the Learner's characteristics, will affect tradeoffs of effort, cost, and time. Stories should not be too complicated—perhaps invoking three causes or less, and no more than six transitions; Klein and Crandall (1995) identified these limits in a study of mental simulation. The building of the explanation will interact with the packaging of the explanation as issues may arise during packaging that will suggest better ways to construct the explanation.

In story building there is a cognitive strain to provide appropriate complexity without being overwhelming. There are several ways to reduce the number of causes so as to increase the Learner's comprehension. One way to maintain the constraint on three causes is to be selective about which causes to include, dropping the ones that are less relevant. Another approach to keep things manageable is to lump several causes—to abstract them into a more general cause.

*Stopping point*. When is the explaining process finished? The stopping point that the Explainer seeks to achieve is for the Learner to experience a perspective shift, as a result of modifying or replacing a belief/assumption, at which point we hypothesize that the Learner should be satisfied. The concept of a perspective shift is different than the "perspective mismatch" we introduced earlier to describe how people may get confused because they make faulty assumptions about other people. Here, in Figure 1, the perspective shift is for the Learner now to appreciate that s/he might well have taken the same actions/decisions as actually occurred given what was known at the time. If the Explainer uses a story, a chain of events in a causal stream, the stopping point is for each transition to appear plausible to the Learner. In the driver/navigator example above, the navigator doesn't bother telling the driver which navigation system is being used, or the logic it relies upon. The navigator simply says "Traffic," because that is sufficient to let the driver appreciate why the surprising direction to turn left was given. The stopping point, the assessment of plausibility, will vary by the Learner's experience and mental models. We are calling this a perspective shift because the Explainer intends for the Learner to move from a mindset that "this isn't making sense," to a mindset of "okay, I see how this all follows." When the Explainer believes the Learner has made this shift, no more discussion is needed and the explanation is completed, having reached a stopping point.

## Global Explaining

Researchers have acknowledged since at least as far back as Clancey (1983), global understanding is an important goal of explanation. In contrast to local explaining, which focuses on what happened during a specific incident, global explaining is about how things work generally—how a device works, how a strategy works, how an organization works. Our model of global explaining was based on the coding analysis, which contained only four global examples, and our general impressions from the larger corpus of global examples. We focused on 19 of those cases because of the completeness of the accounts of "How does x work?" The cases were most often expressed as questions, such as: Why do some modern elevators not stop at the next floor? Why is it so hard to set a digital watch? Why are we often confused by ceiling fans, by airplane reading lights, by the mute button on a TV remote, by motel telephones and clock alarms?

For example, how do ceiling fans work (and why do we sometimes get confused about operating them)? Ceiling fans use a simple interface that doesn't require a monitor or anything fancy—just the cord and the visual of the fan turning.

As Degani (2004) shows, 1 source of confusion is that if the blades are rotating, it is easy to forget how many times you tugged on the cord. And then you won't know if the next tug will increase the rotation speed more or will turn it off. The device does not display its history. All you know is whether or not it is rotating. Further, you don't know how many speed settings it has. Making it more confusing, you don't get instant feedback, as you would with a 3-way bulb. If you are already at the highest speed the next tug will turn it off. But you wouldn't know that because it continues to rotate. So you tug the cord again, starting it up again. Therefore, the essence of understanding why we get confused when operating overhead fans is to grasp how these fans are different from an apparent analog, a three-way light bulb: the delayed feedback that can make it very difficult to control the fan and the difficulty in accurately perceiving the fan speed.

In many ways our account of global explaining is similar to that of local explaining. However, we found two fundamental differences. One difference is that local explanations typically assume that the Learner is familiar with the set-up or with the device, hence the surprise when expectations are violated. Global explanations do not assume familiarity. Therefore, global explanations generally do not focus on the violated expectation.

A second difference is that with local explaining, the Explainer seeks to diagnose the confusion, typically zeroing in on a flaw in the Learner's mental model. The Explainer then seeks to help the Learner revise his/her mental model. For global explaining, the Explainer has no reason to believe that the Learner's mental model is defective and so the Explainer is not seeking to correct the Learner's mental model—only to expand or enrich it, and address some aspect of ignorance. Figure 2 shows a description of global explaining, starting with the issue of how the device or computational system performs a function of interest.

Many of the 42 cases of global explaining in our sample were triggered by curiosity and by a perceived need for more detailed information. The aspects of the Learner's status are essentially the same as in local explaining. The important factors are: richness of the Learner's mental model, the Learner's goals, common ground issues, and time pressure.

We hypothesize that time pressure is less of an issue with global explaining than local explaining, which is triggered by a need to understand and react to a surprise. In addition, the issue of situational understanding does not come up for global explaining because the

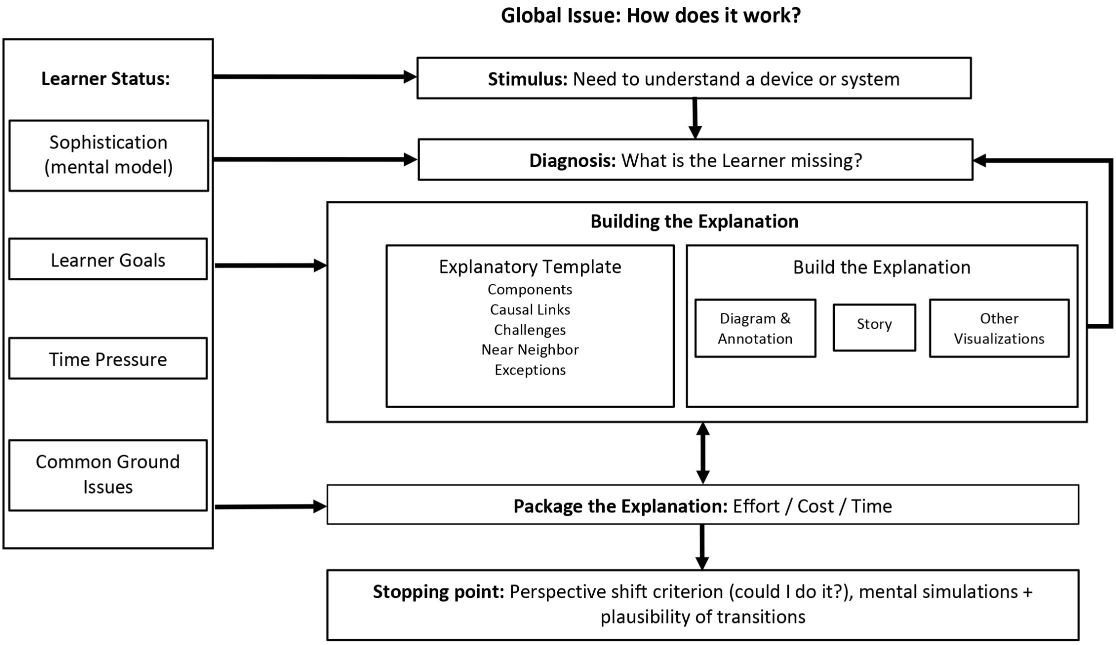
Global Issue: How does it work?
Learner Status:
Sophistication (mental model)
Learner Goals
Time Pressure
Common Ground Issues
Stimulus: Need to understand a device or system
Diagnosis: What is the Learner missing?
Building the Explanation
Explanatory Template
Components
Causal Links
Challenges
Near Neighbor
Exceptions
Build the Explanation
Diagram & Annotation
Story
Other Visualizations
Package the Explanation: Effort / Cost / Time
Stopping point: Perspective shift criterion (could I do it?), mental simulations + plausibility of transitions

*Figure 2.* Model of the process of global explaining.

explaining process is not tethered to a specific incident.

In our sample of 42 cases of global explaining, most appeared to involve some attempt to take the features of the Learner into account and these addressed the first feature, the imagined richness of the Learner's mental model. For example, in explaining why Westerners and Arabs baffle each other in the way they think (Table 1), Klein and Kuperman (2008) identify a small set of cognitive and social mismatches and do not attempt to provide extensive detail on world politics and history; they assume that readers will have some knowledge of the different populations. In explaining why hotel alarm clocks sometimes fail to wake us up (Table 1), Degani (2004) assumed that readers would be familiar with the clock/radios provided by hotels and does not spend time describing their general mechanics, presenting only those details relevant to the traumatic experience of a wake-up failure.

For global explaining, the process of diagnosis is different than for local Explaining. The Explainer is not diagnosing the Learner's confusion or flaws in the Learner's beliefs. Instead, the Diagnosis process in Figure 2 is primarily about the Explainer's speculations about what the Learner is missing. The Learner may be missing a framework if the system is sufficiently strange. Or the Learner may be missing some of the components or some of the links. Or the Learner may be missing causal information that makes the story or the diagram plausible. The Explainer's assessment will guide the way he/she describes the working of the system.

Figure 2 introduces the concept of an Explanatory Template. The template consists of the topics most frequently used in explaining how something works. In reviewing the 19 cases of global explaining in our corpus that we studied in greater depth, we identified several recurring elements. These include: *Components*—The components of the device or computational system. *Links*—The causal links connecting these components. *Challenges*—complications and confusions that warrant a global explanation. *Near neighbors*—There often is a comparable device that can serve as an analog, and the explaining will also describe contrasts with this near neighbor. *Exceptions*—The situations that the device doesn't handle well plus an account of why they are so troublesome, such as the delayed feedback and lack of a history display with the ceiling fan. 13 of the 19 cases included an exception, often as the focus of the explanation. *Tacit knowledge* is often introduced here as the types of knowledge needed to operate the device when it encounters these exceptions.

The Explanatory Template is different than the Causal Palette shown in Figure 1—it is not a compilation of causes frequently invoked to account for events. Rather, it identifies considerations that commonly arise in formulating global descriptions.

Few of the 19 cases included all of these topics. Only AlphaGo and Cruise Control covered all five of the components. Except for the Ceiling Fan case, all but four of these cases included specification of the Components, the Causal Links, the Challenges, the Nearest Neighbor, and the Exceptions. The Ceiling Fan case did not present a specification on Components or Causal Links.

We found that a preferred format for global explaining is a diagram, rather than a story. The need to portray components and causal linkages is better served by a diagram. Further, the diagram format is typically embellished with annotations in order to describe the challenges, the nearest neighbor, the contrasts to that neighbor, and the exceptions.

The features of the Learner's status come into play to select the level of detail the Explainer uses for the elements in the Explanatory Template. The Learner's mental model affects the level of detail most heavily, but the set of goals that motivated the Learner to seek an explanation would also impact the level of detail provided.

What is the stopping point? The Explainer and the Learner are both seeking an outcome in which the Learner can mentally simulate the operation of the device. Each mental simulation is essentially a story, moving from 1 state to another, with plausible transitions. The Learner is trying to imagine how these transitions work. Learners will be satisfied to stop if they feel confident that in most cases they will be able to imagine the system outputs if they are given the system inputs.

Figures 1 and 2 are similar because local and global explaining processes are the same and the high-level structure of the 2 models is the same. Nevertheless, there are some important distinctions in the nature of the components, as noted above. The trigger differs. The nature of the diagnosis differs. The Learner's status is mostly the same but does not involve situational understanding. The nature of the Explanatory Template is very different. The nature of the explanation being built is different—stories for local explaining, diagrams for global explaining.

## DISCUSSION

Rather than attempt to formulate an overall model of the process of explaining we deemed it necessary to develop models for the two types of conditions we observed: local explaining of surprising events and global explaining of the workings of a system or device.

For local explaining the Explainer needs to diagnose the Learner's violated expectancy and then find ways to bring the Learner's expectancies into line. For global explaining, the Explainer needs to help the Learner gain a richer mental model by providing information that the Learner might be missing about the components of the system, their linkages, about the challenges faced by the system developers, about near neighbors to the device, and about exceptions and system failures.

In the Explainable AI community there seems to be a shared belief that to help people understand how a system works the developers need to present them with explanations, and that these explanations need to be accurate, clear, complete, and logical. We argue that none of these properties is clear-cut. Accuracy is a good thing but in complex settings accuracy will depend on the context. An explanation that is accurate under 1 set of conditions may be misleading under other conditions. Clarity is also an attractive virtue until it runs into complexity and dependencies and ambiguities, and the effort to disentangle an explanation will run counter to our desire for a clear, straightforward account. Granted, some complex explanations can be presented in comprehensible ways, but we argue that there is an inherent conflict between adding more details and maintaining comprehension. Completeness seems important until the mass of details renders the explanation incomprehensible. Finally, we value logic, but we usually mean deductive logic rather than inductive or abductive; explanations that violate deductive logic are frowned upon, but in most complex and ambiguous settings deductive logic is not sufficient or even particularly helpful.

Further, we have come to question the value of explanations/statements that are issued without taking the perspective of the Learner into account. We assert that what matters is the process of explaining, which involves an Explainer and a Learner. To be effective, Explainers will need to consider the Learner's background and capabilities—the richness of the Learner's mental model, the Learner's goals, the time pressure the learner is under, potential areas of confusion and common ground breakdown during the dialog, and the way the Learner is assessing the situation. We emphasize the priority of the process of explaining over the issuing of explanations. Our review of the literature had shown that most research was about offering explanations, and that explanations could be assessed on their own merits—their clarity, comprehensiveness, accuracy, logic, predictability—without reflecting the interaction between the Explainer and the Learner.

Complex systems involving AI can benefit from a cognitive engineering analysis of user needs. Such an analysis could help AI developers and systems engineers make discoveries about the causal or explanatory landscape. With this information, developers and systems engineers can determine how best to bring that explanatory information to light given the black-box limitations of machine learning subsets of AI .

Successful explaining should lead to better performance outcomes because the Learner can now do a better job in carrying out a task. One outcome is that the Learner's mental model is elaborated. The Learner has a richer idea of how a system works, how it fails (including how to break it), and how to make it work and manage it (Borders et al., 2019).

We hypothesize that as a result of successful explaining, the Learner should generate more accurate expectancies or predictions about a system or about another person. The Learner should now be able to shift perspectives and see the tasks from the viewpoint of the intelligent system or the other person. The Learner should do better at gauging trust – especially trust in machines, particularly smart machines.

## Conclusions About Local Explaining

*Surprise.* The explaining process is triggered by a surprise, a violated expectation, as opposed to seeing explanations as attempts to fill slots and gaps in knowledge.

*Diagnosis.* Diagnosis is critical on the part of the Explainer to pin down the violated expectation. In contrast, other accounts of explanations start with a complete explication of all the relevant causes and their connections, and view the challenge as trimming details and simplifying—gisting. We disagree. In our view, by diagnosing a single flawed assumption, or a small set of flawed assumptions, the process of explaining is a very focused process rather than trimming details from a comprehensive account.

*Perspective mismatch.* How do we diagnose the reason why another person or a mechanical device made a surprising decision? Our hypothesis is that there is a small set of possible reasons and these can help the Learner make a perspective shift. The reasons we identified are: that person/device might have different knowledge than I do, different goals, they might be operating under different constraints, using different reasoning tactics (which are especially important in dealing with AI), may be aware of different affordances than I am, may have a different mindset, may have sized up the situation differently than I did, or had a different value system than I do.

*Stopping rule.* The stopping rule for explaining is based on a perspective shift in which the Learner gains the ability to see the situation from the vantage point of the other person or the device, so that the "surprising" event is no longer a surprise. For this to happen, the Learner will typically rely on a story of how the surprising event came to pass. If we view the story as a causal chain, the Learner's confidence in the story will depend on a judgment of the plausibility for each of the transitions in a story, or each element and linkage in a diagram. The stopping rule in Figure 1 is subjective—the perspective shift of the Learner and the perception of this perspective shift by the Explainer. Our conclusion is that there are reasonable criteria for cutting off an inquiry without descending into the quest for explanatory depth. And our contribution is that a primary stopping rule is achieving a perspective shift to get to the point such that Learners can appreciate how they might have made the same transitions. At this point, the surprise is no longer surprising.

*Language of reasons.* Explaining relies on a language of reasons. These reasons can be causes, analogs, contrasts, confusions, and stories. The language of reasons, of causality, is different from the language of correlation and the strengthening/weakening of connections between layers in a neural net.

*Contrasts.* Stories often explain by presenting *contrasts*. Our literature review (Mueller et al., 2018) turned up papers asserting that the Learner is not simply wondering why a device recommended course of action x, but rather, why did it recommend x as opposed to y? Our naturalistic study showed that there are other contrasts of interest besides alternative courses of action. There can be contrasts in beliefs, in goals, or in the way the situation is assessed.

## Conclusions About Global Explaining

*Explanatory template.* We postulate an Explanatory Template, which is a set of several items: components of the system, the causal links between the components, the nearest neighbor along with contrasts to that analog, and the exceptions.

*Exceptions.* The last component of the Explanatory Template, the exceptions, is often the richest one for explaining how a system works—or doesn't work. Exceptions provide insight into the inner workings of a program and serve an important function in reminding us that Machine Learning systems rely on very different reasoning strategies than people do. Our position stands in contrast to many accounts that view the goal of explanations as building a mental model of how a device works. Our view is that the Explainer also needs to describe how the device does not work—its limitations—along with workarounds to handle these limitations.

*Diagrams.* Global explaining typically depends on a diagram of internal structure, often with annotations, as opposed to the story format for local explaining.

*Stopping point.* The stopping point for global explaining is to get the Learner to be able to run a mental simulation with standard starting conditions and be reasonably confident in the output of the mental simulation.

## Limitations of the Research

This project was designed as an initial investigation into the process of explaining, using qualitative methodology. We cast a wide net in collecting the cases, but we also relied heavily on one particular source, Degani (2004), which may have affected our analyses. Therefore, the models we developed and the conclusions we present must be treated as tentative. Other researchers, using different methods or even different sets of materials, may arrive at somewhat different accounts, but we anticipate that the core notions in our results and analyses will replicate. On the other hand, we expect that efforts to apply the two models of explaining, local and global, will reveal the shortcomings of the models presented in Figures 1 and 2 and will result in better accounts than ours for understanding human explaining and for improving the explainability of AI systems.

## General Conclusions and Recommendations

The first 10 years of research on Explainable AI produced AI-oriented or computer scientist-oriented explanations, and did not consider the needs of users who are not computer scientists (Mueller et al., 2018). These researchers in the field of Explainable AI had access to knowledge structures that we now regard as interpretable (rules, goals, decision trees, etc.). Unsurprisingly, these early efforts still failed. The few scholars who looked at this carefully in the following decades (Brézillon, 1994; Brézillon & Pomerol, 1997; Clancey, 1987; Doyle et al., 2003; Kass & Finin, 1988; Sørmo et al., 2005; Swartout & Moore, 1993) all seemed to identify the problem as relating to the explanation needs of humans. This includes understanding the context, goals, knowledge and the like of the user, and having a user model. You have to know what the person's goal is and what their knowledge is to provide a reasonable explanation.

Swartout, Clancey, and others realized that "interpretable" to computer science professionals means "justificational," not explanatory. Computer scientists in the present XAI community seem to echo the error of the past (i.e., that a rule trace is an explanation, or what Swartout and Moore called the "Myth of recap-as-explanation"). Further, "interpretable" has a particular and formal model-theoretic meaning to computer scientists, which is quite distinct from the everyday, psychological meaning.

Our recommendation is to learn from the failures of the past, and orient explanation to the needs of the person, not the information in the system.

Many XAI researchers acknowledge that explanations need to help the user or Learner develop a better mental model of the AI system. This better mental model should help the user make better predictions about the system, understand when it is out-of-bounds, develop better expectations, develop workarounds, and so on. The Learner needs to make judgments about appropriate trust and also needs to understand the boundary conditions of the device.

The issue of mental models cuts in both directions. For interaction with an AI system, one outcome of effective explaining is that the Learner form a better mental model of the system, which is an important and achievable goal. However, our work suggests that the Explainer can be more successful by having a good mental model of the Learner, and this remains a challenge for AI systems and an opportunity for future generations of AI systems. We acknowledge that the use cases illustrated in the present article do not show how a system, AI or otherwise, might actually accomplish this in the way that a human explainer might, nor do we offer any recommendations for identifying the richness of a Learner's mental model to improve the practice/development of AI. Nevertheless, we suggest that AI developers might make progress by appreciating the way perspective-taking, diagnosis, and other psychological aspects of explaining come into play.

We also intend for our models of local and global explaining to go beyond AI and to have application to person-to-person interactions. Issues of perspective taking and common ground are central to dialog and to effective coordination.

## ACKNOWLEDGMENTS

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. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of AFRL, the Department of Defense, or the U.S. Government.

## SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

## REFERENCES

Borders, J., Klein, G., & Besuijen, R. (2019). *An operational account of mental models: A pilot study* [Conference session]. Proceedings of the 2019 International Conference on Naturalistic Decision Making, San Francisco, CA.

Brézillon, P. (1994, *September). Context needs in cooperative building of explanations* [Conference session]. First European Conference on Cognitive Science in Industry, pp. 443–450.

Brézillon, P., & Pomerol, J. C. (1997, *April). Joint cognitive systems, cooperative systems and decision support systems: A cooperation in context* [Conference session]. European Conference on Cognitive Science, Manchester, UK, pp. 129–139.

Bureau of Enquiry and Analysis for Civil Aviation Safety (BEA). (2012). *Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-ZCP operated by Air France flight AF 447 Rio de Janeiro - Paris*. Ministere de l'Ecologie, du Developpement durable, des transpors et du Logement, Paris, France.

Clancey, W. J. (1983). The epistemology of a rule-based expert system—a framework for explanation. *Artificial Intelligence*, *20*(3), 215–251. https://doi.org/10.1016/0004-3702(83)90008-5

Clancey, W. J. (1987). *Knowledge-based tutoring: The GUIDON program*. MIT press.

Collins, H. M. (1992). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Degani, A. (2004). *Taming HAL: Designing interfaces beyond 2001*. Palgrave/Macmillan.

Doyle, D., Tsymbal, S., & Cunningham, P. (2003). *Department of Computer Science, Trinity College*. Dublin.

Forbus, K. D., & Feltovich, P. J. (Eds.). (2001). *Smart machines in education*. AAAI Press.

Goguen, J. A., Weiner, J. L., & Linde, C. (1983). Reasoning and natural explanation. *International Journal of Man-Machine Studies*, *19*(6), 521–559. https://doi.org/10.1016/S0020-7373(83)80070-4

Hoffman, R., Klein, G., & Miller, J. (2011). Naturalistic investigations and models of reasoning about complex indeterminate causation. *Information Knowledge Systems Management*, *10*(1-4), 397–425. https://doi.org/10.3233/IKS-2012-0203

Hoffman, R. R., Mueller, S. T., & Klein, G. (2017). Explaining explanation, Part 2: Empirical foundations. *IEEE Intelligent Systems*, *32*(4), 78–86. https://doi.org/10.1109/MIS.2017.3121544

Kass, R., & Finin, T. (1988). The need for user models in generating expert system explanations. *International Journal Of Expert Systems*, *1*(4).

Klein, G. (2018). Explaining explanation, Part 3: The causal landscape. *IEEE Intelligent Systems*, *33*(2), 83–88. https://doi.org/10.1109/MIS.2018.022441353

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, *21*(5), 88–92. https://doi.org/10.1109/MIS.2006.100

Klein, G., Calderwood, R., & Clinton-Cirocco, A. (2010). Rapid decision making on the fire ground: The original study plus a postscript. *Journal of Cognitive Engineering and Decision Making*, *4*(3), 186–209. https://doi.org/10.1518/155534310X12844000801203

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. In W. B. Rouse & K. R. Boff (Eds.), *Organizational simulation*. Wiley.

Klein, G., & Jarosz, A. (2011). A naturalistic study of insight. *Journal of Cognitive Engineering and Decision Making*, *5*(4), 335–351. https://doi.org/10.1177/1555343411427013

Klein, G. A., & Crandall, B. W. (1995). The role of mental simulation in naturalistic decision making. In J. Flach, P. Hancock, J. Caird, & K. Vicente (Eds.), *The ecology of human-machine systems* (pp. 324–358). Lawrence Erlbaum Associates.

Klein, H. A., & Kuperman, G. (2008). Through an Arab cultural lens. *Military Review*, *88*(3), 100.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *6*(8), 539–551. https://doi.org/10.1111/j.1747-9991.2011.00413.x

McKeown, K. R., & Swartout, W. R. (1987). Language generation and explanation. *Annual Review of Computer Science*, *2*(1), 401–449. https://doi.org/10.1146/annurev.cs.02.060187.002153

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2018). "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:190201876.

Palmer, B. (2013). *Understanding Air France 447*. William Palmer, Jr.

Pearl, J. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Swartout, W. R., & Moore, J. D. (1993). Explanation in second generation expert systems. *Second Generation Expert Systems*, *543*, 585.

Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, *24*(2), 109–143. https://doi.org/10.1007/s10462-005-4607-7

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning*. Macmillan.

Gary Klein, Ph.D., is a cognitive psychologist who helped to initiate the Naturalistic Decision Making movement in 1989. His Recognition-Primed Decision (RPD) model has been tested and replicated several times. He also developed a Data/Frame model of sensemaking, and a Triple Path model of insight. His work relies on Cognitive Task Analysis methods that he and his team developed. He has formulated the Pre-Mortem method for identifying risks and the ShadowBox method for training cognitive skills. He founded Klein Associates in 1977 and sold it to Applied Research Associates in 2005. He started his new company, ShadowBox LLC, in 2014.

Dr. Hoffman is a recognized world leader in cognitive systems engineering and Human-Centered Computing. He is a Senior Member of the Association for the Advancement of Artificial Intelligence, Senior Member of the Institute of Electrical and Electronics and Engineers, Fellow of the Association for Psychological Science, Fellow of the Human Factors and Ergonomics Society, and a Fulbright Scholar. His Ph.D. is in experimental psychology from the University of Cincinnati. His Postdoctoral Associateship was at the Center for Research on Human Learning at the University of Minnesota. He served on the faculty of the Institute for Advanced Psychological Studies at Adelphi University. He has been Principal Investigator, Co- Principal Investigator, Principal Scientist, Senior Research Scientist, or Principal Author on over 60 grants and contracts including alliances of university and private sector partners. He has been a consultant to numerous government organizations. He has been recognized internationally in the fields of psychology, remote sensing, human factors engineering, intelligence analysis, weather forecasting, and artificial intelligence—for his research on the psychology of expertise, the methodology of cognitive task analysis, human-centering issues for intelligent systems technology, and the design of macrocognitive work systems. Hiscurrent work focuses on “Explainable AI.”

Dr. Shane T. Mueller is an associate professor in the Department of Cognitive and Learning Sciences, and is also affiliated with the Department of Computer Science at Michigan Technological University. He studies applied human performance and cognition via both empirical research and the development of computational and mathematical approaches, with a focus on measurement of human and Artificial Intelligence performance and behavior. His research has been funded by a number of agencies, including DARPA, in which he was supported by the BICA program to develop the BICA Cognitive Decathlon (a comprehensive plan for testing biologically-inspired artificial intelligence across a broad spectrum of skill categories) and the XAI program to develop measurement approaches and psychological theories of explanation. His research as also been funded by IARPA, DTRA, AFRL, the Ford Foundation, ICANN, and others. He is also the developer of the PEBL Test Battery (), a software testing platform for measuring human cognitive skill across a wide range of neuropsychological tests.

Emily Newsome, B.S., is an Assistant Manager and Research Associate with ShadowBox LLC. In her work with ShadowBox, she has developed scenarios and conducted cognitive interviews with nurses, child welfare workers, and military personnel. She received her Bachelor’s degree in Psychology from Wayne State University (Detroit, MI) in 2014.