Explanation to Avert Surprise

Melinda Gervasio, Karen Myers, Eric Yeh, Boone Adkins
SRI International
333 Ravenswood Avenue, Menlo Park, California 94025, USA
{firstname.lastname}@sri.com

ABSTRACT

Most explanation schemes are reactive and informational: explanations are provided in response to specific user queries and focus on making the system's reasoning more transparent. In mixed autonomy settings that involve teams of humans and autonomous agents, *proactive explanation* that anticipates and preempts potential surprises can be particularly valuable. By providing timely, succinct, and context-sensitive explanations, autonomous agents can avoid perceived faulty behavior and the consequent erosion of trust, enabling more fluid collaboration. We present an explanation framework based on the notion of *explanation drivers*—i.e., the intent or purpose behind agent explanations. We focus on explanations meant to *reconcile* expectation violations and enumerate a set of *triggers* for proactive explanation. Most work on explainable AI focuses on intelligibility; investigating explanation in mixed autonomy settings helps illuminate other important explainability issues such as purpose, timing, and impact.

Author Keywords

Explainable autonomy; explainable AI; human-machine teams; collaborative AI; intelligibility

ACM Classification Keywords

H.5.m. Information interfaces and presentation: Miscellaneous; I.2.m Artificial intelligence: Miscellaneous

INTRODUCTION

Humans judge mistakes by computer systems more harshly than mistakes by other humans, with errors having a disproportionately large impact on perceived reliability [1,2]. This negative impact on trust has particularly significant repercussions for human-machine teams, where the humans' trust in the autonomous agents directly affects how well they utilize the agents. The effect is particularly unfortunate when the human perceives an agent to be misbehaving when in fact it is behaving appropriately but in response to conditions unknown to the user.

We propose that a primary motivation for explanation should be *surprise*. When an agent violates expectations—typically, not in a good way—a human collaborator will invariably want to know the reason why. Reacting to the human's surprise and explaining away the violation is a valid approach, but even more effective would be if the agent could *anticipate* the surprise and *proactively explain* what it is about to do. This averts a potentially unpleasant surprise that distracts the user and erodes trust.

Our foray into explainable autonomy began a few years ago, when we were developing autonomous agents for a project on team autonomy for uncertain, dynamic, adversarial environments in mixed human-machine settings. As we observed the agents in action, we would sometimes see puzzling behavior—for example, an agent might suddenly change course away from its intended destination. Our first thought would almost invariably be that there was a problem with the agent but, upon further inspection, we would realize that the agent had good reason for its action. For example, it might be reacting to an unexpected event or diverting to a higher-priority task. A straightforward UI showing the agents' current plans helped somewhat to alleviate this problem, but this was a solution targeted at the autonomous agents' designers, not at the end users who would be teaming with these automated agents in the future.

The need for intelligent systems that could explain themselves was recognized early on with expert systems [8], with the desire of both system developers and end users to better understand the reasoning behind a computational system's conclusions to determine whether it could be trusted. More recently, the dominant work in explanation has been on explaining the decisions of learned classifiers, particularly in the context of interactive learning [6,9], recommender systems [3,10], and deep learning [5,11].

Explanation for autonomy differs in a number of ways. The decision to be explained is typically part of a larger, orchestrated sequence of actions to achieve some long-term goal. Decisions occur at different levels of granularity—from overarching policies and strategic decisions down to individual actions. Explanation is required for various reasons under different circumstances: before execution to explain planning decisions, during execution to explain deviations from planned or expected behavior, and afterwards to review agent actions.

In the collaborative human-machine team settings that we are primarily interested in, whether humans serve as supervisors or as teammates, explanation during execution presents the additional challenge of limited cognitive resources. With the human already engaged in a cognitively demanding task, system explanations must be succinct, timely, and context-sensitive. In particular, when a human asks, "Why are you doing that?" it will often be because the

ExSS '18, March 11, Tokyo, Japan.

agent has done something unexpected and the agent's explanation must address that.

EXPLANATION DRIVERS

We have developed an explanation framework based on the concept of *explanation drivers*: the intent or purpose behind an agent's explanation. We distinguish between three classes of drivers: *Inform, Reconcile,* and *Prime*.

Explanations to *Inform* are what most people typically think of as explanations. They provide straightforward answers to basic *wh*-questions—for example, "What is your goal?" or "How do you plan to achieve that goal?" or "Where are you going?" In the mixed-team setting, *Inform* explanations are particularly useful early on, when the human is still trying to get an overall sense of an (unknown) agent's decision-making. However, even after some level of trust has been already been established, *Inform* explanations often still remain useful for maintaining that trust.

Explanations that *Reconcile* address expectation violations. They answer questions borne of surprise—e.g., "What are you doing!" or "Why aren't you doing X?" or "Why did you do Y [instead of Z]?" *Reconcile* explanations are most effective when presented before the consequences of the decision are apparent, to prevent the surprise in the first place. For example, a warning from a firefighting drone that it will be diverting to help extinguish a fire that is growing faster than expected avoids surprising the user and possibly causing concern. It also gives the user the opportunity to change the plan—for example, to send the drone to its original target and to co-opt a different one to help instead.

Finally, there are explanations that *Prime* the user for assistance. Just as in human teams, communication and coordination is critical in mixed teams. In human-supervised settings, an important part of this collaboration involves agents recognizing when they need help and providing humans with the information they need to provide appropriate guidance. Beyond simply asking for help, *Prime* explanations inform humans why help is needed to help them provide appropriate assistance. For example, if the agent has low confidence in its best action, it can let the human supervisor confirm or override.

Here we focus on *Reconcile* explanations—in particular, on *proactive* explanations designed to avoid unpleasant surprises for human collaborators. This decision to focus on proactive explanations was partially validated by the results of a small four-person user study we conducted in mid-2017. The study was in a fictional domain of drone firefighting and rescue, and participants were given the task of understanding what the drones were doing, with the knowledge that world was dynamic (e.g., fires could start and die out on their own) and that all information was uncertain (e.g., groups to be rescued could appear and disappear, fires could be larger/smaller than expected). Participants were presented with snapshots of an evolving scenario. In the baseline condition, they were provided with basic information about current drone assignments and the status of all known fires and groups, and they could ask basic questions about the drones' behavior. In the proactive condition, they were also given preemptive explanations (as textual pop-ups) of certain drone decisions.

The participants all found the proactive explanations to be useful. As one participant put it, "[Proactive explanations were] very helpful, particularly anything that was counterintuitive or represented a big change." Based on the questions participants asked, we observed that everyone wanted to know the big picture, both in terms of the overall plan and the agents' overall priorities. In addition, the participants expected the drones to address all the targets—fires extinguished and groups rescued—with a strong preference for saving people over extinguishing fires.

TRIGGERS FOR PROACTIVE EXPLANATION

Most explanation schemes are reactive: explanations about system decisions are generated on-demand in response to specific user queries. While reactive explanations are useful in many situations, proactive explanations are sometimes called for, particularly in mixed autonomy settings where, for example, close coordination is required and humans are engaged in tasks of their own or are supervising large teams. Proactive explanations serve to keep the human's mental model of the agent's decisions aligned with the agent's actual decision process, minimizing surprises that can distract from and disrupt the team's activities. Used judiciously, they can also reduce the communication burden on the human, who will have less cause to question the agent's decisions.

We propose the use of *surprise* as the primary motivation for proactivity, with agents using potential expectation violations to trigger explanation. Identifying expectation violations requires having a model of the user's expectations. However, instead of relying on a comprehensive formal model of the human's expectations based on a representation of team and individual goals, communication patterns, etc., we identify classes of expectations based on the simpler idea of expectation norms. That is, given a cooperative team setting where the humans and the agents have the same objectives, we set out to determine expectations on agent behavior based on rational or commonsense reasoning. We enumerate a set of triggers for proactive explanation, discussing for each one the manifestation of surprise, the expectation violation underlying the surprise, and the information that the proactive explanation should impart (Table 1). The triggers are not an exhaustive list but include a broad range that we have found particularly useful in our work on explainable autonomy.

Lim & Dey's investigation of intelligibility demands is focused on context-aware applications [7]; however, some of their findings regarding the situations in which different explanations apply are relevant here. In particular, inappropriate actions, critical situations, situations

Trigger	Surprise	Expectation	Explanation
Historical deviations	Action differs from past behavior in similar situations	Agent will behave as it has in the past	Acknowledgement of unexpected action
Unusual situations	Atypical action	Normal operation	Information about unusual situation
Human knowledge limitations	(Seemingly) incorrect action	Agent has the same information as the human	Indicate decision criteria
Preference violations	Non-preferred action	Agent will adhere to specified preferences	Acknowledgment of violation with rationale
Indistinguishable effects	Different action	Agent will perform 'obvious' action	Information about equivalent options
Plan deviations	Action contrary to plan	Actions according to plan	Change of plans and rationale
Indirect trajectories	(Seemingly) aimless behavior	Agent will move toward goal	Plan for getting to goal

Table 1. Triggers for proactive explanation and their surprise manifestations, underlying causes, and explanation content.

involving user goals, and high external dependencies were all found to increase the need for intelligibility, particularly through *why not* and situation explanations.

Historical Deviations

An important aspect of trust is predictability—a human will generally expect an agent to perform the same actions that it has performed in similar situations in the past. Thus, an agent suddenly executing a different action is likely to surprise the user. An agent can anticipate this situation through a combination of statistical analysis of performance logs and semantic models for situation similarity. Explanation involves an acknowledgment of the atypical behavior and the rationale behind it—for example, "Aborting rescue mission because of engine fire."

Unusual Situations

A human observer lacking detailed understanding of a domain may be aware of actions for normal operations but not of actions for more unusual situations. An agent's actions in these situations may thus surprise the user. The agent can identify these situations by their frequency of occurrence—for example, if the conditions that triggered the behavior are below some probability threshold. Explanation to avert this type of surprise involves describing the unusual situation to the user. For example, an agent might explain, "Normal operation is not to extinguish fires with civilians on board but fire is preventing egress of Drone 17 with a high-priority evacuation."

Human Knowledge Limitations

Sensing and computational capabilities, particularly in distributed settings, can enable autonomous platforms to have insights and knowledge that are unavailable to human collaborators. Through awareness of decisions based on this information, an agent can identify potential mismatches in situational understanding that can lead to surprising the user with seemingly incorrect decision-making. Explanation involves identifying the potential mismatch and surfacing that to the user. For example, Google Maps already does this to some extent when it suggests an unusual route along with the justification that it is currently the best option given current traffic conditions.

Preference Violations

Many formulations of autonomy incorporate preferences over desired behaviors, whether created by the system modeler at design time or imposed by a human supervisor later on. When making decisions, an agent will seek to satisfy these preferences; however, various factors (e.g., resource limitations, deadlines, physical restrictions) may require that they be violated, leading to the agent seemingly operating contrary to plan and surprising the user. Explanation in this case involves acknowledging the violated preference or directive and providing the reason why—for example, "Entering no-fly zone to avoid dangerously high winds."

Indistinguishability of Effects

Two actions may be very different in practice but achieve similar effects—for example, different routes of similar duration to the same destination. This can surprise a human observer who may not have realized their comparable effects or even been aware of the other (chosen) option. Agents can anticipate this type of surprise by measuring the similarity of actions or trajectories and of outcomes. Explanation then involves making the human aware of different options with similar impact—for example, "I will extinguish Fire 47 before Fire 32 but extinguishing Fire 32 before Fire 47 would be just as effective."

Plan Deviations

Agents are expected to be executing a plan to achieve a goal. Inevitably, situations will arise that require a change of plans which, if initiated by the agent, can cause surprise. Absent an explicit shared understanding of the current plan, an agent can rely on an expectation of inertia—that is, that it will continue moving in the same direction, toward the same target. By characterizing this tendency and recognizing (significant) changes, the agent can anticipate potential surprise. Explanation involves informing the user of the plan change—for example, “Diverting to rescue newly detected group.” This may be sufficient if it calls attention to a new goal or target previously unknown to the user but if the change involves a reprioritization of existing goals, explanation also needs to include the rationale—for example, “Diverting to rescue Group 5 before Group 4 because fire near Group 5 is growing faster than expected.”

Indirect Trajectories

More generally, agents are expected to be engaged in purposeful behavior. In spatiotemporal domains, observers can typically infer from an agent’s trajectory its destination and, based on that, its goal. For example, a drone heading toward a fire is likely to be planning to extinguish the fire. Surprises occur when the agent has to take an indirect route and appears to be headed nowhere meaningful. The agent can identify this situation by determining the difference between its actual destination and an apparent one, if any. Explanation then involves explicitly identifying the goal and the reason for the indirect action—for example, “New task to retrieve equipment from supply depot.”

SUMMARY AND CONCLUSIONS

Prior work has noted the utility of surprise for driving intelligent system behavior. Recognizing that the most valuable information to users is information that complements what they already know, Horvitz et al. [4] focus on surprising predictions as the situations about which to alert the users in a traffic forecasting system. Wilson et al. [12] use surprise in an intelligent assistant for software engineering to entice users to discover and utilize programming assertions. Here, we use surprise to drive proactive explanations and help users understand decisions that might otherwise cause concern.

We are currently investigating our approach to proactive explanation in various explainable autonomy formulations. In one where an autonomous controller selects, instantiates, and executes plays from a pre-determined mission playbook, we identify surprising role allocations based on assignment to suboptimal resources and use degree of suboptimality to drive proactivity. In another involving a reinforcement learner acquiring policies in a gridworld domain, we use sensitivity analyses that perturb an existing trajectory to identify points where relatively small changes in action lead to very different outcomes.

Focusing on the motivation behind explanations in collaborative autonomy settings helps bring to light issues not often addressed in work on explainable AI. We present a framework for explanation drivers, focusing in particular on explanations for reconciling expectation violations. We argue that averting surprise should be a primary motivation for explanation and enumerate a set of triggers for proactive explanations. While most current work on explanation focuses opaque deep learning models and is thus primarily concerned with interpretability, mixed autonomy settings require additional metrics to capture the usefulness and significance of explanations in terms of their quality and impact. Ultimately, our objective is to provide evidence that explanations enable the appropriate and effective use of intelligent agents in mixed autonomy settings.

REFERENCES

1. Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2014. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experimental Psychology: General* 144, 1.
2. Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *Int. J. Human-Computer Studies* 58: 697–718.
3. Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. *Proc CSCW 2000.*
4. Eric Horvitz, Johnson Apacible, Raman Sarin, and Lin Liao. 2005. Prediction, expectation, and surprise: methods, designs, and study of a deployed traffic forecasting service. *Proc. UAI 2005*.
5. Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. *Proc. CVPR 2015*.
6. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. *Proc. IUI 2015*.
7. Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. *Proc. UBICOMP 2009.*
8. Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green, and Stanley N. Cohen. 1975. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research* 8, 4: 303–320.
9. Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *Int. J. Human-Computer Studies* 67(8): 639–662.
10. Nava Tintarev and Judith Masthoff. 2007. Effective explanations of recommendations: user-centered design. *Proc. RecSys’07*.
11. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why should I trust you?” Explaining the predictions of any classifier. *Proc. KDD 2016.*
12. Aaron Wilson, Margaret Burnett, Laura Beckwith, Orion Granatir, Ledah Casburn, Curtis Cook, Mike Durham, and Greg Rothermel. 2003. Harnessing curiosity to increase correctness in end-user programming. *Proc. CHI 2003*.