What Should Be in an XAI Explanation? What IFT Reveals

Jonathan Dodge, Sean Penney, Andrew Anderson, Margaret Burnett
Oregon State University
Corvallis, OR; USA
{ dodgej, penneys, anderan2, burnett }@eecs.oregonstate.edu

ABSTRACT

This workshop's call for participation poses the question: *What should be in an explanation?* One route toward answering this question is to turn to theories of how humans try to obtain information they seek. Information Foraging Theory (IFT) is one such theory. In this paper, we present lessons we have learned about how IFT informs Explainable Artificial Intelligence (XAI), and also what XAI contributes back to IFT.

CCS Concepts

•**Human-centered computing** → **User studies;** •**Computing methodologies** → *Intelligent agents;*

Author Keywords

Intelligent Agents; Explainable AI; Intelligibility; Content Analysis; Video Games; StarCraft; Information Foraging

INTRODUCTION

Explainable AI (XAI) is burgeoning to help ordinary users understand their intelligent agents' behavior – but many fundamental questions remain in order to achieve this goal. This paper describes our recent progress toward one such question: *What should be in an explanation?*

We have been working to answer this question in a domain often used for AI research, namely Real-Time Strategy (RTS) games, from two sides. First, to understand what a high-quality *supply* of explanations might contain, we conducted a qualitative analysis of the utterances of expert explainers [2]. Second, to understand *demand* for explanations in the same domain, we conducted a user study [10] to understand the questions participants formulated when assessing an intelligent agent playing the popular RTS game StarCraft II [9]. Here, we focus on the latter study.

There have been previous explorations into what should be in an XAI explanation [1, 6, 7, 8, 14, 15], but few such explorations draw upon theories of how humans problem-solve. We used Information Foraging Theory (IFT) [12] to help fill this gap and approach our investigation. IFT is based on a predator-prey model [12]. Grounded in prior work about how people seek information [3, 11], we used StarCraft II to investigate how both the expert explainers (suppliers) and our participants ("demanders") would navigate the information environment as they sought to make sense of a game while it unfolded.

In the RTS domain, players compete for control of territory by fighting for it. Each player raises an army to fight their opponents, which takes resources and leads players to build *Expansions* (new bases) to gain more resources. Players also can use resources to create *Scouting* units, which lets them learn about their enemies' movements to enable *Fighting* in a strategic way. For a more in-depth explanation of the domain, refer to [9].

In our user study [10] investigating this domain, we gave 20 experienced StarCraft II players a game replay file[1] to analyze and asked them to record whatever they thought were *key decision points* (i.e., any "event which is critically important to the outcome of the game") during the match. Participants worked in pairs, allowing us to keep them talking by leveraging the social convention of conversing about their collaborative task. Because we wanted to understand how the participants go about assessing an intelligent agent's decisions, we told them that one of the players in the game was under AI control. However, this was not true; both players were human professionals.

[1]We used game 3 of this match (http://lotv.spawningtool.com/23979/) from the *IEM Season XI - Gyeonggi* tournament.

Figure 1. A screenshot from our study, with participants anonymized (bottom right corner). Superimposed red boxes point out: (1, bottom left) the *Minimap*, a birds-eye view enabling participants navigate around the game map; (2: top left) a drop-down menu to display the *Production tab* for a summary of build actions in progress; (3, middle right) *Time Controls* to rewind/forward or change speed.

ExSS 2018, March 11, 2018, Tokyo, Japan.

The participants' main task was to assess the AI's capabilities. To do so, the participants replayed the game using the built-in StarCraft tool, shown in Figure 1, which offers the ability to observe the previously recorded events. The tool provided functionality to freely navigate with the camera, pause/rewind with time controls, and drill down into various aspects of the game state, helping participants decide how the AI was doing.

After participants finished the main task, we conducted a retrospective interview in two parts. In both parts, we asked participants questions about things they had said and done, while pointing them out in the video we had just made of those participants working on the task. In the first part, we navigated to each decision point they identified and asked why it was so important. In the second part, we asked about selected navigations using questions based on previous work [11], such as "*What about that point in time made you stop there?*". A more detailed methodology can be found in [10].

WHAT WE'VE LEARNED SO FAR: IFT → XAI

Things we've learned from studying Prey

In IFT, predators seek *prey*, which are the pieces of information in the environment they they think they need. In the context of XAI, such prey are evidence of the agents decision process, which are then used to create explanations for agents' actions.

To investigate the information participants were trying to obtain, we analyzed the questions that they asked each other. We categorized their questions according to the Lim-Dey intelligibility types [7], which separate questions into `What`, `What-Could-Happen`, `Why-Did`, `Why-Didn't`, and `How-To`. We also added a `Judgment` intelligibility type to capture when participants sought a quality judgment.

Although most previous XAI research has found `Why` to be highly demanded information, our participants rarely sought `Why` or `Why-Didn't` information. Instead, our participants showed a strong preference for asking `What` questions.

What was so interesting about `What`? The participants' `What` information seeking was about finding out more about state than they currently knew. Our participants did so primarily in three categories: drill down, higher level, and temporal.

Drill down `What`s usually involved participants spatially navigating around the map, sometimes opening up objects or menus to access more detailed game state information. E.g. *"Is the human building any new stuff now?"* The second category, higher-level `What`s, involved trying to abstract a little above the details, to gain a higher level of understanding of the game state. E.g. *"What's going on over there?"* The third category, temporal `What`s, involved finding out more about differences or similarities in state over time. E.g. *"When did he start building...?"*

Finally, to investigate whether our distribution of `What` vs. `Why` results were reasonably representative for this domain, we compared our participants' questioning (i.e., explanation demand) against the answers (explanation supply) produced by professional explainers in this domain, namely shoutcasters [2]. The results showed that the shoutcasters' commentaries in StarCraft games [2] matched well with the above explanation demands. In particular, shoutcasters' utterances were mostly about the `What` intelligibility type, with very few utterances of the `Why` or `Why-Didn't` types. Further, the shoutcasters were remarkably consistent with each other in frequency of using each intelligibility type. The consistency between the supply-side and demand-side results offers evidence that in the RTS domain, `What` explanation content is more in demand than `Why` or `Why-Didn't`.

Implications: Taken together, these results show that in this domain, participants placed *very* high value on state information — but not always at the same granularity, and not always restricted to a single moment in time. How an XAI system can satisfy these explanation needs may not be straightforward, but one of our findings suggests a way forward: Shoutcasters may be usable as a gold standard. That is, the remarkable similarity between the frequency of shoutcasters' utterances (supply) and participants' desired prey (demand) for most intelligibility types suggests that XAI explanation systems in the RTS domain may be able to model their explanation content, timing, and construction, around shoutcasters' explanations.

Things we've learned from studying Paths

In IFT, prey exists within some *patch(es)*, and the forager navigates between patches by following *paths*, made up of one or more *links*. Investigating the paths participants used revealed a great deal of information about the kinds of costs they can incur in the RTS domain when seeking information.

Traditionally, IFT looks at the navigation cost to get to a patch (here, explanation), usually in number of clicks, and the cognitive cost of absorbing the necessary information in the patch once there. These costs are relevant to XAI too, but our investigation discovered participants incurred significant cognitive effort in both path discovery and path triage.

Why so expensive? Professional RTS players perform several hundred actions per minute (APM), and each such action potentially destroys or updates the available foraging paths. This produces an information environment in which foraging paths are numerous, rapidly updating, and have limited lifespan. Thus, our participants were faced with many more potentially useful foraging paths than they could possibly follow, and had to spend significant effort just *choosing* a path.

Some coped with these costs by adhering to a single foraging path throughout the task, rewinding rarely. These participants minimized their cognitive costs of choosing, but paid a high *information cost*, because by not following other paths, they missed out on potentially explanatory information. Others chose not to pay this information cost, and instead paid a *navigation cost* by often rewinding and pausing to spatially explore. Rewinding also incurs substantial *cognitive cost*, as more context information must be tracked — but that extra context may provide useful explanatory power.

[2] Shoutcasters are sportscasters for e-sports. They perform a similar sort of analysis as our participants were doing, but with the added constraint that they must analyze and explain the game in real-time, so they cannot pause or rewind.

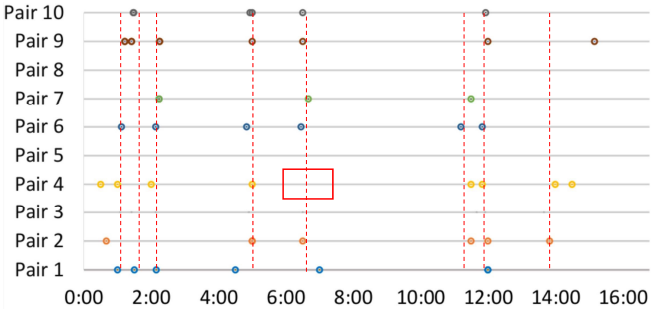

Figure 2. Dots show the *Building-Expansion* decision points each participant pair (y-axis) identified over time (x-axis). Red lines show when *Expansion* events actually occurred. (Participants noticed most of them.) The red box shows where Pair 4 failed to notice an event they likely wanted to note, based on their previous and subsequent behavior.

Interestingly, when costs of choice were low, participants' explanation seeking followed fairly traditional foraging patterns. For example, early in the game, participants scrutinized the game objects carefully and in detail — a sharp contrast to late in the game when many more game objects and foraging paths were present. This could suggest that as the information environment grows in complexity, users in this domain will seek explanations at a higher level of abstraction (i.e., a group of units as opposed to a single unit).

Implications: XAI explanation system would benefit from incorporation of an explanation recommender. Such a recommender could take into account both the human cognitive cost of considering too many paths when few can be followed, and the information cost of neglecting some path too long. For example, if the domain is well known to an explanation system a priori, such a recommender may help guide users (reducing their cognitive cost of choosing) to the explanations that are the most important (reducing the information cost of missing important explanatory information). In this case, it appears we know *Expansions* are important before any analysis occurs.

Things we've learned from studying Scent and Cues

Recall that we requested that participants write down key decision points. To forage for these, they followed *cues*, which are information features connected to links in the environment. In our study, cues were the same across sessions, because everyone replayed the same game. Unlike cues, *scent* is "in the head" – it is foragers' assessment of cues' meaning.

Unfortunately, participants missed information that we suspect they would have found key, because some cues distracted them. Our videos showed that what participants were looking at when they were distracted — the "distractor cues" — tended to be combat-oriented and affected even simple game states.

For example, in Figure 2, a nearly full *decision column* suggests that participants tended to agree that this decision was key. A nearly full *participant pair row* suggests that this pair consistently found this *type* of decision to be key. Thus, missing dots (e.g., see the red box) correspond to times when a participant pair was distracted from a key event.

In fact, the scents emanating from some types of cues seemed to overpower others consistently. Consider the example in Figure 3, which shows how *Fighting* tended to overpower

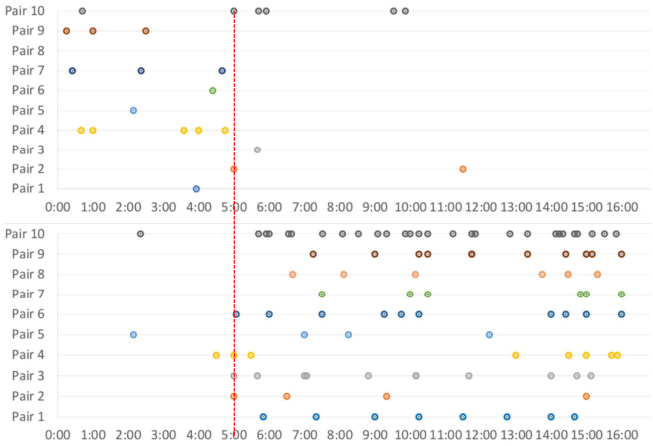

Figure 3. (Top:) The *Scouting* decision points identified by our participant pairs (y-axis), with game time on the x-axis. (Bottom:) The *Fighting* decision points identified, plotted on the same axes. After *Fighting* events begin (red line), *Scouting* decision points are no longer noticed often — despite important *Scouting* actions continuing to occur.

Scouting. The top image in the figure shows all of the *Scouting* decision points our participants identified, while the bottom image shows all of the *Fighting* decision points. The red line going through both images is the point at which combat first begins – and also the time when scouting is usually last noticed, despite being ongoing throughout the game.

Implications: Distractions abound, and may be systematic. Some facets of the environment may elicit emotional response and receive undue attention as a result. In this case, participants preferred investigating *Fighting* over *Scouting*.

Another implication relates to an XAI explanation system's user interface's support for human workflow. Paths are easily forgotten in the presence of interruptions. Each new action may interrupt the current foraging path, which leads to people forgetting things – made worse by sheer path quantity. Previous research has found that To-Do Listing [5] is an effective strategy to help prevent users from forgetting so much.

WHAT WE'VE LEARNED SO FAR: XAI → IFT

In the previous sections, we focused on things we learned about XAI by applying IFT to our data set. Now, we turn the other direction, since this study is the first to apply IFT to XAI and the RTS domain. The RTS domain presents an extremely complex and rapidly changing environment, more so than other IFT environments in the literature like Integrated Development Environments (IDEs) and web sites [3, 4, 11, 13]. In the RTS domain, hundreds of actions happen *each minute*. Further, the environment is continually affected by actions which do not originate from the forager.

As discussed, our participants were faced with many paths, and had to rapidly triage which paths to follow. This presents an interesting IFT challenge. Previous research [11] identified a "scaling up problem" in IFT — a difficulty estimating value/cost of distant prey as the path to the prey became long. In our case, we observed that foraging paths were short, but since so many paths are available, not much time is available to make an accurate path value/cost estimate. The current study reveals a "breadth version" of this scaling problem (Figure 4).

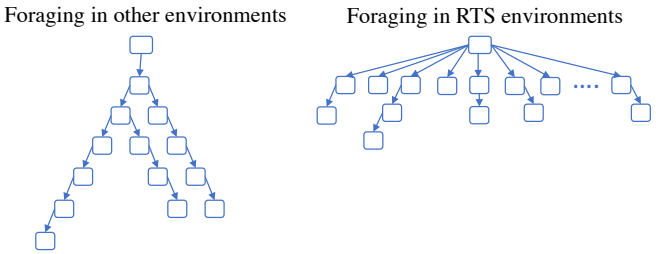

Figure 4. Conceptual drawing of foraging in the RTS domain vs. previously studied foraging. (Left): Information environments in prior IFT literature: the predator considers few paths, but the paths are sometimes very deep. This figure inspired by an IDE foraging situation in [[11] Fig. 5]. (Right): Foraging in RTS, where most navigation paths are shallow, but with numerous paths to choose from at the top level.

Turning to prey, in the XAI setting, the prey is evidence of the agent's decision process. Establishing trust in an XAI system requires the user to know how it behaves in many circumstances. Thus, the prey is "in pieces" – meaning that bits of it are scattered over many patches. As previous work [11] has shown, "prey in pieces" creates foraging challenges, because finding and assembling all the bits can be tedious and error-prone. In the model-agnostic XAI setting, IFT's "prey in pieces" problem becomes even more pronounced, because of the uncertain relationships between causes/effects, or even whether the agent will ever behave the same way again.

CONCLUSION

This paper summarizes the first investigation into information foraging behaviors shown by participants tasked with assessing a RTS intelligent agent. Our formative studies used IFT to inform XAI and vice versa, by examining both *supply* (expert explanations) and *demand* (user's questions).

Our use of the IFT lens allows us to leverage results obtained from applying IFT to non-XAI domains, while also improving the ability to transport/generalize findings among XAI domains. By connecting XAI to IFT foundations, we can bring to XAI a real theoretical foundation based on what informations humans want and how they look for it.

ACKNOWLEDGMENTS

This work was supported by DARPA #N66001-17-2-4030 and NSF #1314384. Any opinions, findings and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of NSF, DARPA, the Army Research Office, or the US government.

REFERENCES

1. S. Amershi, M. Cakmak, W. Knox, and T. Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
2. J. Dodge, S. Penney, C. Hilderbrand, A. Anderson, L. Simpson, and M. Burnett. 2018. How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games. In *ACM Conference on Human Factors in Computing Systems*. To Appear.
3. S. Fleming, C. Scaffidi, D. Piorkowski, M. Burnett, R. Bellamy, J. Lawrance, and I. Kwan. 2013. An information foraging theory perspective on tools for debugging, refactoring, and reuse tasks. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 22, 2 (2013), 14.
4. W. Fu and P. Pirolli. 2007. SNIF-ACT: A cognitive model of user navigation on the world wide web. *Human-Computer Interaction* 22, 4 (2007), 355–412.
5. V. Grigoreanu, M. Burnett, and G. Robertson. 2010. A strategy-centric approach to the design of end-user debugging tools. In *ACM Conference on Human Factors in Computing Systems*. ACM, 713–722.
6. T. Kulesza, M. Burnett, W. Wong, and S. Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *ACM International Conference on Intelligent User Interfaces*. ACM, 126–137.
7. B. Lim and A. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *ACM International Conference on Ubiquitous Computing*. ACM, 195–204.
8. B. Lim, A. Dey, and D. Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *ACM Conference on Human Factors in Computing Systems*. ACM, 2119–2128.
9. S. Ontañón, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, and M. Preuss. 2013. A Survey of Real-Time Strategy Game AI Research and Competition in StarCraft. *IEEE Transactions on Computational Intelligence and AI in Games* 5, 4 (2013), 293–311.
10. S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, L. Simpson, and M. Burnett. 2018. Toward Foraging for Understanding of StarCraft Agents: An Empirical Study. In *ACM Conference on Intelligent User Interfaces*. To Appear.
11. D. Piorkowski, A. Henley, T. Nabi, S. Fleming, C. Scaffidi, and M. Burnett. 2016. Foraging and navigations, fundamentally: developers' predictions of value and cost. In *2016 ACM International Symposium on Foundations of Software Engineering*. ACM, 97–108.
12. P. Pirolli. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford Univ. Press.
13. S.S. Ragavan, S. Kuttal, C. Hill, A. Sarma, D. Piorkowski, and M. Burnett. 2016. Foraging among an overabundance of similar variants. In *ACM Conference on Human Factors in Computing Systems*. ACM, 3509–3521.
14. S. Stumpf, E. Sullivan, E. Fitzhenry, I. Oberst, W. Wong, and M. Burnett. 2008. Integrating rich user feedback into intelligent user interfaces. In *ACM International Conference on Intelligent User Interfaces*. ACM, 50–59.
15. J. Vermeulen, G. Vanderhulst, K. Luyten, and K. Coninx. 2010. PervasiveCrystal: Asking and answering why and why not questions about pervasive computing applications. In *Intelligent Environments (IE), 2010 IEEE International Conference on*. IEEE, 271–276.