Horses For Courses: Making The Case For Persuasive Engagement In Smart Systems

Simone Stumpf
Centre for HCI Design
City, University of London
London, UK
Simone.Stumpf.1@city.ac.uk

ABSTRACT

Current thrusts in explainable AI (XAI) have focused on using interpretability or explanatory debugging as frameworks for developing explanations. We argue that for some systems a different paradigm – persuasive engagement – needs to be adopted, in order to affect trust and user satisfaction. In this paper, we will briefly provide an overview of the current approaches to explain smart systems and their scope of application. We then introduce the theoretical basis for persuasive engagement, and show through a use case how explanations might be generated. We then discuss future work that might shed more light on how to best explain different kinds of smart systems.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models** • Human-centered computing → Interaction design theory, concepts and paradigms

KEYWORDS

Explanations; explanatory debugging; intelligibility; smart systems, intelligent user interfaces; argumentation.

ACM Reference format:

Simone Stumpf. 2019. Horses For Courses: Making The Case For Persuasive Engagement In Smart Systems. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 6 pages.*

INTRODUCTION

Explainable AI (XAI) has gained attention in recent years, with significant research efforts being expended to investigate how to generate interpretable explanations [4][30][12], how to manage and structure the explanation design process [6][36], and the principles and important concepts underlying various approaches to provide explanations for smart systems [14][20] in order to increase user satisfaction [36][9], user trust and/or reliance [5], decrease misuse or disuse [25], make users' mental models more sound [15], or more deeply involve the user in interactive machine learning, human-in-the-loop learning, and decision-making [14][38][13][10]. A long-standing focus of research in XAI has been what and how to explain to users of AI [34][26][32][20][18][35][27], both in terms of content e.g. data, details of the algorithm used, etc. and presentation e.g. textual, graphical, visualizations, etc.

However, there is increasing evidence that explanations might have differing and even conflicting effects on users [3], and that they have to be carefully crafted to the context in which explanations are provided [31][33]. This position paper reviews existing XAI frameworks which currently shape the design and deployment of explanations. We will show that these frameworks have underlying assumptions that make them unsuitable for all situations. Instead, designers and developers would do well to consider the purpose and intended effects of explanations that are provided, in order to inform the content and presentation. We will introduce persuasive engagement as an alternative framework for shaping explanations, and provide a use case that shows how explanations arise from this framework. We close by discussing the road ahead for work in XAI and potential future work investigating the persuasive engagement framework.

EXISTING EXPLANATION FRAMEWORKS

There are currently two main frameworks that shape how researchers shape explanations: interpretability (sometimes also called intelligibility or transparency) and explanatory debugging. We will provide a brief overview of each of these frameworks and show that they are making various assumptions that shape in which contexts they might be usefully deployed. Table 1 shows an overview of the main differences of these two frameworks.

IUI Workshops'19, March 20, 2019, Los Angeles, USA.

Table 1. Main differences between interpretability and explanatory debugging frameworks

Aspect	Interpretability	Explanatory Debugging
Context of Use	Incompleteness of AI system in optimization or evaluation	Interactive machine learning, personalization
Main Goals	Interpretability, users' understanding	Correct system "bugs"
Secondary Goals	Fairness, reliability, trust	Users' understanding, satisfaction
Explanation design – What to include	Explanations types, such as What, Certainty, Why, Why Not and Inputs	Interactive explanations including features, predictions, and model (e.g. weights, prediction confidence, class balance)
Explanation design – How to present	Communicate in "human-understandable" terms	Presented iteratively, as sound and complete as possible while not overwhelming the user

The Interpretability Framework

Interpretability [4] applies to machine learning systems that have "the ability to explain or to present in understandable terms to a human." It has been argued that only those systems in which incompleteness arises in optimization or evaluation require an explanation; systems which do not have "significant consequences for unacceptable results" or which are "common-place" will not need an explanation [4]. Once an AI system is interpretable, other desirable aspects of AI systems, such as fairness, reliability, trust, will also follow along.

Aligned with this framework is work developed in the context of context-aware and pervasive systems [19], that sense, learn and adapt themselves to their environment and users. In this context, a number of explanations types, such as What, Certainty, Why, Why Not and Inputs have been identified which should be presented to users in order to increase interpretability. Explanations are judged on their quality when compared to human explanations, and thus the main thrust of research in this framework is to find generic dimensions of interpretability that could lead to quality being optimized, such as how well patterns in data or reasons for specific decisions are communicated, how easily biases and errors are identified, and how much user information processing constraints are taken into account. Working within this framework, research efforts have concentrated on how best to expose the workings of AI algorithms to its users, either through making algorithms more interpretable (e.g. [12]) or investigating ways in which patterns, data, biases, etc. could be communicated to users (e.g. [35][11]).

The Explanatory Debugging Framework

A different approach to providing explanations is the framework of explanatory debugging [14]. Key aims in this framework are to help the user identify the "bugs" in machine learning and communicate enough of the machine learning system so that the user can make targeted and useful changes to improve the system to address these bugs. Explanations are provided to users in order to build better mental models of how the intelligent system behaves to support interactive machine learning. Ideally, this is also associated with increased user satisfaction if system performance improves, for example by the system personalizing itself to user preferences, or making better decisions but this is only a corollary to the main aim of improving system performance. Research has suggested that explanations should be presented iteratively and be as sound and complete as possible while not overwhelming the user; the user feedback should be able to incrementally modify the system behavior in a meaningful way while also being reversible [14]. Particular ways to expose the logic of these systems in aid of interactive machine learning have been investigated, including how to allow the user to interact with the explanations interactively to provide feedback to the system [32][16][8][1][17].

PERSUASIVE ENGAGEMENT

We are not suggesting that one of these frameworks is better than the other; in fact, we argue that the choice of framework is dependent on the context and purpose in which explanations are to be deployed. There is not one right explanation framework and instead we need to consider the best 'horses for courses'. There is some evidence [2] that not all applications need an explanation which would accord with the interpretability framework. We argue that both frameworks do not serve smart systems well that sit outside of their scope: those that might have "inconsequential" effects, those that are common-place but need to gain the trust of users, or ones that do not learn from user interactions. For example, many smart heating systems do not have "significant consequences for unacceptable results"; all you do is change the heating setting. Siri's and Alexa's mistakes provide for much hilarity and viral Internet memes but rarely do we want to turn to interpretability or explanatory debugging frameworks for creating explanations for them. Eiband et al's [6] work on a fitness app also does not fall nicely with either of these existing frameworks. Yet, users (and industry developing these applications) want explanations for these kinds of systems, especially if they go wrong.

We have previously argued [33] that these kinds of systems need to be compatible with constrained engagement [38], where the user can engage with the system to input their preferences or override the system if necessary but communication from the system is constrained so it does not overwhelm the user or push itself to the front. The main aim of the communications between system and the user is to increase user trust and satisfaction. Explanations in these situations about system decisions need to be as concise and light-weight as possible and do not need to be as detailed as in the interpretability and explanatory debugging frameworks that hope to increase the understanding of users. We argue that to help shape explanations for these kinds of applications and situations, a framework of persuasive engagement might be helpful. Table 2 outlines the main aspects of persuasive engagement. This framework draws heavily on previous seminal work in argumentation and rhetoric, which will be outlined next.

Table 2. Main aspects of persuasive engagement framework

Aspect	Persuasive Engagement
Context of Use	Everyday low-risk systems, constrained engagement situations
Main Goals	User trust and satisfaction
Secondary Goals	Understanding
Explanation design – What to include	Inputs, Inference step, Decision/Behavior
Explanation design – How to present	Concise, lightweight, drill-down on demand

Argumentation and Rhetoric

Previous work in AI using argumentation approaches [29][24] [23][22] has mainly focused on how to represent and reason about decisions, to generate explanations automatically using arguments for and against a decisions, or how to draw on inference categories to enrich the persuasiveness of explanations. In contrast, our work uses argumentation and rhetoric to provide guidance about what to include in an explanation, and possibly how to present it.

Argumentation has been significantly influenced by the work of Toulmin [37] who proposed that an argument has the structure shown in Figure 1. The most basic form of an argument is data (also known as premises or facts) and its link to a qualified conclusion (i.e. the conclusion could be more or less certain), and it usually suffices because it draws on accepted inference steps for the targeted audience. An argument is thus a set of premises that support a conclusion with some degree of plausibility; an explanation contains arguments for and against the conclusion, often without needing to give the actual details of the inference step [22]. Rhetorical argumentation is concerned with "increasing the adherence of a particular audience" to a conclusion [28] and therefore focuses its research on what inference steps are persuasive for certain audiences [28].

In the argument structure proposed by [37] further 'why' questions by the person the argument is directed at might trigger additional elements to be provided: warrants and backing provide further reasons as to why the inference is valid, whereas rebuttals might be drawn out that affect the certainty of the conclusion.

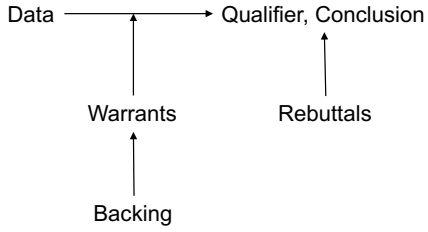

Figure 1. Toulmin argumentation scheme

Application to a Smart Heating Use Case

We now present how to generate an explanation within the persuasive engagement framework, drawing reference to the argument structure presented in the previous section (Table 3).

Table 3. Mapping between persuasive engagement and argument structure

Persuasive engagement	Argument structure
Inputs	Data/Facts
Persuasive reason for making the decision	Inference step
Decision/Behavior	Qualified Conclusion
On request: show input values	On 'Why': Show Warrants, Backing, Rebuttals
Present in easily understandable form	Natural language

To generate an explanation for a decision (i.e. the conclusion) in this framework, we simply expose the inputs (i.e. the data) that are used to make the decision and the reason for making the decision or behavior (i.e. inference step). Only if the user requests more information, does the explanation provide further, more detailed input values (i.e. warrants, backing, and rebuttals). The inference step draws on reasons that the intended user group finds "agreeable" or persuasive, and thus might change depending on the targeted user group. Ideally, these explanations are in a form that the intended user will easily understand, such as text, or simple graphics or visualization, etc.

We now present a worked use case in smart heating systems for persuasive engagement. Our research investigated increasing trust and understanding of the smart heating system, specifically the app that allowed users to manage and control their heating (Figure 2).

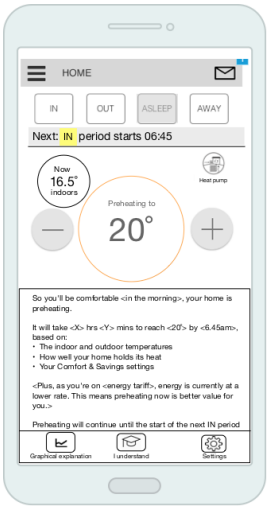

Figure 2. Overview of control app

Our program of work was set in a UK project to understand the overall value and user experience of hybrid heat pump deployment in demand-response settings. This project, FREEDOM[1] (Flexible REsidential Energy Demand Optimisation and Management), led by Passiv Systems Ltd. and funded by Western Power Distribution and Wales and West Utilities.

[1] https://www.westernpower.co.uk/projects/freedom

Our endeavor sought to explain system behavior through transparency design [6]. The results of a previous user study [31] indicated that the system needed to provide explanations to users when unexpected behavior occurred, at the point of or even prior to starting this behavior. We also found that textual explanations and simple visualizations were preferred by users, and they wanted reasons for system behavior that included reference to their comfort and cost.

To start, we collaborated with an expert heating engineer employed by our collaborative partner, Passiv Systems Ltd, to generate a list of all system behaviors. These were when the system decides to:

- pre-heat the home to reach a temperature setpoint for a period in the user's schedule when they have indicated that they will be at home;
- heat to maintain a temperature setpoint if the user is at home;
- not heat and run at a lower temperature than the setpoint if the user is at home;
- heat at a higher temperature than the one set for when the user is at home;
- switch between heat sources;
- to not heat when the user is not at home or asleep;
- to implement demand-response (i.e. shift the heating pattern due to network demand, and variable energy tariffs)

For each of these behaviors, we then needed to explain their respective inputs (i.e. the data that is drawn on to make a conclusion). For each of the 7 decisions, we again interviewed the heating expert from Passiv Systems Ltd. to investigate the inputs that the algorithm used to make each of the above decisions. Pre-heating was one of the most complex behaviors in terms of inputs and also one of the most misunderstood system behaviors in a previous study [31]; all other decisions used considerably less variety of inputs. We therefore illustrate the design of how to explain using this rich example. For pre-heating the home, we found that the following inputs mattered:

- Current internal temperature;
- Current external temperature;
- Learnt properties around the rate of heating of the home;
- Schedule and associated temperature setpoint;
- User preference to optimize comfort versus cost;
- 24-hour weather forecast;
- Tariff information for heat sources.

Once we had all of this information, we began to iteratively design and prototype the presentation of explanations for all these behaviors, following the persuasive engagement framework. In its simplest form, a textual explanation included at least one statement explaining the reasons (i.e. inference step) that gave a motivation for linking the data to the decision, and a set of inputs that underlie the decision. For example, in the textual explanation for pre-heating (Figure 3), we included the overall motivation for preheating (Figure 3 A). We drew on comfort reasons by drawing attention to "so you are comfortable in the <morning>". In addition, if a demand-response situation arises where tariffs increase based on network demand, we add an additional reason about reducing energy costs: "Plus, [...] this means pre-heating now is better value for you." The interface also lists the key components and data that the behavior was based on, as shown in Figure 3 B.

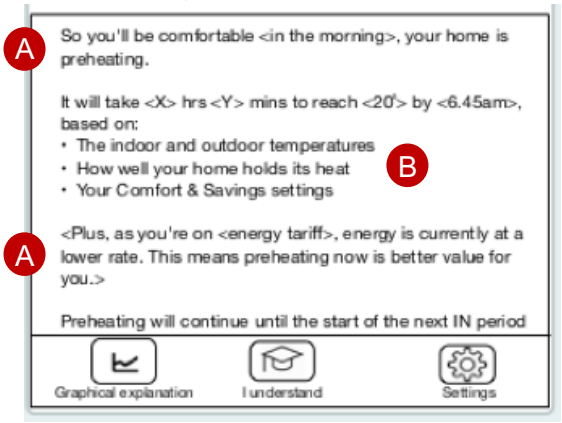

Figure 3. Textual explanation for pre-heating.

The user can switch to a graphical explanation on request by pressing a control at the bottom of the screen, thereby indicating that they want to have a deeper explanation of the heating system's behavior, akin to asking a further 'why' question. A graphical explanation visualizes the main inputs underlying the system behavior with their concrete values. Each timeline shows the current time in the middle of the x-axis. The left part of the graph shows the input values up to the current time, on the right is a projected forecast of what the system will do in the future, shown partially transparent and in dashed lines, to indicate uncertainty. For example, in pre-heating (Figure 4) a wide variety of data determines preheating to reach an indoor temperature. It depicts the current schedule, the current outdoors temperature, the tariff information (in case of demand response situations), and the current trade-off setpoints for comfort versus savings. It shows the period of time when the system has been or will be pre-heating to achieve the set temperature points when people are expected to be in the home.

Discussion and Future Work

We have described the current main frameworks in existence and their scope of application. We have introduced a new framework – persuasive engagement – by drawing on argumentation theory, and shown how it might be applied in a use case. Our view fits well with how explanations are seen in the social sciences as information about causality and counterfactuals in answer to a 'why' question [21]. In addition, it comes closest to what is termed a "pragmatic view" of explanations [7]. We firmly believe that any advances in XAI need to involve inter-disciplinary efforts to contribute new thoughts and research directions.

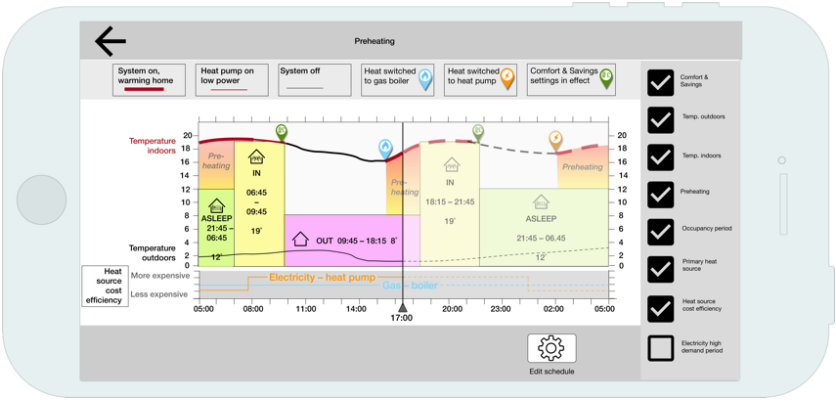

Figure 4. Detailed wireframe for graphical explanation of pre-heating.

Our work has three main avenues for future work: 1) its scope and validation, 2) the components of explanations in this framework, 3) new presentation styles for persuasive engagement.

First, we need to consider when it would be appropriate to apply persuasive engagement. We have shown this in one use case but we need to investigate the boundaries of when this framework needs to be abandoned and instead interpretability or explanatory debugging become more appropriate or useful. Hand in hand goes validation of how robust this framework is, or whether we are missing important aspects. This will necessarily take a longer-term approach in which researchers and practitioners pool their experiences. It will also require that we are explicit about which kinds of systems we are addressing and for which purposes.

Second, we need to investigate each component of the framework more deeply. For example, what are the user groups that we might want to target with persuasive engagement? What inference steps are acceptable for certain user groups? How do we determine the inputs that matter? What process should we adopt in this case? In the case of smart heating systems, we involved both users and experts in transparency design [6] to determine inference steps and the inputs, but these details obviously are dependent on the domain and targeted user group. Hence more work is needed that relate to how we should construct the 'What' of an explanation.

Last, presentation is another area that needs to be investigated. Currently, we are focusing on textual versus graphical or visualizations, but are there other, novel forms of presentations that we should consider? In rhetoric, affect and emotions matter – how much do they matter in explanations following persuasive engagement? Also important is the question of scale, and when explanations crafted through persuasive engagement might become overwhelming.

In summary, the research effort to study XAI is far from over, and instead our work to investigate making AI understandable, trustworthy and effective has just begun.

ACKNOWLEDGMENTS

This work was supported by the FREEDOM project, funded by Western Power Distribution, and Wales and West Utilities. We thank Graeme Aymer from City, University of London, and Tom Veli, Edwin Carter, Frasier Harding and Tim Cooper from Passiv Systems Ltd. for their help with this research.

REFERENCES

[1] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2010. Examining multiple potential models in end-user interactive concept learning. In *Proceedings of the 28th international conference on Human factors in computing systems*, 1357–1360. https://doi.org/10.1145/1753326.1753531

[2] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (IUI '12), 169–178. https://doi.org/10.1145/2166966.2166996

[3] A. Bussone, S. Stumpf, and D. O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, 160–169. https://doi.org/10.1109/ICHI.2015.26

[4] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[5] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6: 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[6] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces* (IUI '18), 211–223. https://doi.org/10.1145/3172944.3172961

[7] Eiband, Malin, Schneider, Hanna, and Buschek, Daniel. Normative vs. Pragmatic: Two Perspectives on the Design of Explanations in Intelligent Systems. In *Joint Proceedings of the ACM IUI 2018 Workshops co-located with the 23rd ACM Conference on Intelligent User Interfaces (ACM IUI 2018).* https://doi.org/http://ceur-ws.org/Vol-2068/exss7.pdf

[8] A. Groce, T. Kulesza, Chaoqiang Zhang, S. Shamasunder, M. Burnett, Weng-Keen Wong, S. Stumpf, S. Das, A. Shinsel, F. Bice, and K. McIntosh. 2014. You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems. *IEEE Transactions on Software Engineering* 40, 3: 307–323. https://doi.org/10.1109/TSE.2013.59

[9] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 241–250. https://doi.org/10.1145/358916.358995

[10] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning* 95, 3: 423–469. https://doi.org/10.1007/s10994-013-5413-0

[11] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the 28th international conference on Human factors in computing systems*, 1343–1352. https://doi.org/10.1145/1753326.1753529

[12] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*, 2668–2677. Retrieved December 11, 2018 from http://proceedings.mlr.press/v80/kim18d.html

[13] Todd Kulesza, Margaret Burnett, Simone Stumpf, Weng-Keen Wong, Shubhomoy Das, Alex Groce, Amber Shinsel, Forrest Bice, and Kevin McIntosh. 2011. Where Are My Intelligent Assistant's Mistakes? A Systematic Testing Approach. In *End-User Development*, Maria Francesca Costabile, Yvonne Dittrich, Gerhard Fischer and Antonio Piccinno (eds.). Springer Berlin Heidelberg, 171–186. Retrieved December 16, 2013 from http://link.springer.com/chapter/10.1007/978-3-642-21530-8_14

[14] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 126–137. https://doi.org/10.1145/2678025.2701399

[15] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (CHI '12), 1–10. https://doi.org/10.1145/2207676.2207678

[16] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In *Proceedings of the 2010 IEEE Symposium on Visual Languages and Human-Centric Computing* (VLHCC '10), 41–48. https://doi.org/10.1109/VLHCC.2010.15

[17] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented End-user Debugging of Naive Bayes Text Classification. *ACM Trans. Interact. Intell. Syst.* 1, 1: 2:1–2:31. https://doi.org/10.1145/2030365.2030367

[18] Brian Y. Lim and Anind K. Dey. 2010. Toolkit to Support Intelligibility in Context-aware Applications. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (UbiComp '10), 13–22. https://doi.org/10.1145/1864349.1864353

[19] Brian Y. Lim and Anind K. Dey. 2011. Investigating Intelligibility for Uncertain Context-aware Applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (UbiComp '11), 415–424. https://doi.org/10.1145/2030112.2030168

[20] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 27th international conference on Human factors in computing systems* (CHI '09), 2119–2128. https://doi.org/10.1145/1518701.1519023

[21] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]*. Retrieved from http://arxiv.org/abs/1706.07269

[22] Martin Možina. 2018. Arguments in Interactive Machine Learning. *Informatica* 42, 1. Retrieved December 14, 2018 from http://www.informatica.si/ojs-2.4.3/index.php/informatica/article/view/2231

[23] Martin Možina, Jure Žabkar, and Ivan Bratko. 2007. Argument based machine learning. *Artificial Intelligence* 171, 10: 922–937. https://doi.org/10.1016/j.artint.2007.04.007

[24] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-Based Explanations in Recommender Systems: Conceptual Framework and Empirical Results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (UMAP '18), 293–298. https://doi.org/10.1145/3213586.3225240

[25] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2: 230–253. https://doi.org/10.1518/001872097778543886

[26] Michael J. Pazzani. 2000. Representation of electronic mail filtering profiles: a user study. In *IUI*, 202–206. https://doi.org/10.1145/325737.325843

[27] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. 2018. Toward Foraging for Understanding of StarCraft Agents: An Empirical Study. In *23rd International Conference on Intelligent User Interfaces* (IUI '18), 225–237. https://doi.org/10.1145/3172944.3172946

[28] Chaim Perelman and Louise Olbrechts-Tyteca. 1971. *The New Rhetoric: a treatise on Argumentation*. University of Notre Dame Press.

[29] Iyad Rahwan and Guillermo R. Simari. 2009. *Argumentation in artificial intelligence*. Springer.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16), 1135–1144. https://doi.org/10.1145/2939672.2939778

[31] Simonas Skrebe and Simone Stumpf. 2017. An exploratory study to design constrained engagement in smart heating systems. In *Proceedings of the 31st British Human Computer Interaction Conference*.

[32] 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. Hum.-Comput. Stud.* 67, 8: 639–662.

[33] Simone Stumpf, Simonas Skrebe, Aymer, Graeme, and Hobson, Julie. 2018. Explaining Smart Heating Systems to Discourage Fiddling with Optimized Behavior. In *Joint Proceedings of the ACM IUI 2018 Workshops co-located with the 23rd ACM Conference on Intelligent User Interfaces (ACM IUI 2018)*. https://doi.org/http://ceur-ws.org/Vol-2068/exss13.pdf

[34] William R. Swartout. 1983. XPLAIN: a system for creating and explaining expert consulting programs. *Artif. Intell.* 21, 3: 285–325. https://doi.org/10.1016/S0004-3702(83)80014-9

[35] J. Talbot, B. Lee, A. Kapoor, and D. S. Tan. 2009. EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the 27th international conference on Human factors in computing systems*, 1283–1292.

[36] Nava Tintarev and Judith Masthoff. 2007. Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on Recommender systems*, 153–156. https://doi.org/10.1145/1297231.1297259

[37] Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge,UK.

[38] Rayoung Yang and Mark W. Newman. 2013. Learning from a Learning Thermostat: Lessons for Intelligent Systems for the Home. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp '13), 93–102. https://doi.org/10.1145/2493432.2493489