Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda

Ashraf Abdul[1], Jo Vermeulen[2], Danding Wang[1], Brian Y. Lim[1], Mohan Kankanhalli[1]
[1] School of Computing, National University of Singapore, Singapore
[2] Department of Computer Science, Aarhus University, Denmark
ashrafabdul@u.nus.edu · jo.vermeulen@cs.au.dk · wangdanding@u.nus.edu
brianlim@comp.nus.edu.sg · mohan@comp.nus.edu.sg

ABSTRACT
Advances in artificial intelligence, sensors and big data management have far-reaching societal impacts. As these systems augment our everyday lives, it becomes increasingly important for people to understand them and remain in control. We investigate how HCI researchers can help to develop accountable systems by performing a literature analysis of 289 core papers on explanations and explainable systems, as well as 12,412 citing papers. Using topic modeling, co-occurrence and network analysis, we mapped the research space from diverse domains, such as algorithmic accountability, interpretable machine learning, context-awareness, cognitive psychology, and software learnability. We reveal fading and burgeoning trends in explainable systems, and identify domains that are closely connected or mostly isolated. The time is ripe for the HCI community to ensure that the powerful new autonomous systems have intelligible interfaces built-in. From our results, we propose several implications and directions for future research towards this goal.

Author Keywords
Intelligibility; explanations; explainable artificial intelligence; interpretable machine learning.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Artificial Intelligence (AI) and Machine Learning (ML) algorithms process sensor data from our devices and support advanced features of various services that we use every day. With recent advances in machine learning, digital technology is increasingly integrating automation through algorithmic decision-making. Yet, there is a fundamental challenge in balancing these powerful capabilities provided by machine learning with designing technology that people feel empowered by. To achieve this, people should be able to *understand* how the technology may affect them, *trust* it and feel in *control*. Indeed, prior work has identified issues people encounter when this is not the case (e.g., with smart thermostats [183] and smart homes [131, 184]). Algorithmic decision-making can also affect people when they are not directly interacting with an interface. Algorithms are used by stakeholders to assist in decision-making in domains such as urban planning, disease diagnosis, predicting insurance risk or risk of committing future crimes, and may be biased (e.g., [167, 185]).

To address these problems, machine learning algorithms need to be able to *explain* how they arrive at their decisions. There has been increased attention into interpretable, fair, accountable and transparent algorithms [38, 146] in the AI and ML communities, with examples such as DARPA's Explainable AI (XAI) initiative [74] and the "human-interpretable machine learning" community. Recently, the European Union approved a data protection law [57, 65] that includes a "right to explanation", and USACM released a statement on algorithmic transparency and accountability [171]. The time is clearly ripe for researchers to confront the challenge of designing transparent technology head on.

However, much work in AI and ML communities tends to suffer from a lack of usability, practical interpretability and efficacy on real users [50, 100, 132]. Given HCI's focus on technology that benefits *people*, we, as a community, should take the lead to ensure that new intelligent systems are transparent from the ground up. This is echoed by Shneiderman *et al*. [158], who discussed the need for interfaces that allow users "to better understand underlying computational processes" and give users "the potential to better control their (the algorithms') actions" as one of the grand challenges for HCI researchers. Several researchers are already contributing towards this goal, e.g. with research on interacting with machine-learning systems [1, 90, 100, 101, 162, 183], algorithmic fairness [107, 108] and accountability [43, 47].

As a first step towards defining an HCI research agenda for explainable systems, this paper maps the broad landscape around explainable systems research and identifies opportunities for HCI researchers to help develop autonomous and intelligent systems that are explainable by design. We make three contributions:

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.

ACM ISBN 978-1-4503-5620-6/18/04...$15.00.
https://doi.org/10.1145/3173574.3174156

- Based on a literature analysis of 289 core papers and 12,412 citing papers, we provide an overview of research from diverse domains relevant to explainable systems, such as algorithmic accountability, interpretable machine learning, context-awareness, cognitive psychology, and software learnability.
- We reveal fading and burgeoning trends in explainable systems and identify domains that are closely connected or mostly isolated.
- We propose several implications and directions for future research in HCI towards achieving this goal.

RELATED WORK

We grouped prior work into three areas: related landscape articles relevant to explainable artificial intelligence, work on intelligibility and interpretability in HCI, and methods to analyze trends in a research topic. For brevity and to foreshadow the results of our literature analysis, we will highlight only a few key research areas.

Explainable Artificial Intelligence (XAI) Systems

There has been a surge of interest in explainable artificial intelligence (XAI) in recent years driven by DARPA's initiative to fund XAI [74]. Historically, there has been occasional interest in explanations of intelligent systems over the past decades with expert systems in the 1970s [165, 172], Bayesian networks (for a review, refer to [102]) and artificial neural networks [6] in the 1980s, and recommender systems in the 2000s [33, 81]. The recent successes of AI and machine learning for many highly visible applications and the use of increasingly complex and non-transparent algorithms, such as deep learning, calls for another wave of interest for the need to better understand these systems.

The response from the AI and ML communities has been strong with a wide range of workshops: Explanation-aware Computing (ExaCt) [150] at ECAI 2008, 2010-2012 and AAAI Symposia 2005, 2007. Fairness, Accountability, and Transparency (FAT-ML) workshop at KDD 2014-2017 [9], ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI) at [93], NIPS 2016 Interpretable ML for Complex Systems [181], IJCAI 2017 Workshop on Explainable AI [1]. Workshops have also been organized at HCI venues: CHI 2017 Designing for Uncertainty in HCI [69], CHI 2016 Human-Centred Machine Learning [61], and IUI 2018 Explainable Smart Systems [118].

This has produced a myriad of algorithmic and mathematical methods to explain the inner workings of machine learning models; see [15] for a survey. However, despite their mathematical rigor, these works suffer from a lack of usability, practical interpretability and efficacy on real users. For instance, Lipton [119] argues that there is no clear agreement on what interpretability means, and provides a taxonomy of both the reasons for interpretability and the methods to achieve interpretability. Doshi-Velez and Kim [50] attempted to better define what *interpretability* means, and how one can measure whether a system is interpretable. They provide an overview of methods for interpretability evaluation and discuss open challenges. Others have attempted to map research regarding intelligibility, explanations or interpretable algorithms. For example, Miller [132] charts an overview of research in the social sciences (philosophy, psychology, and cognitive science) regarding how people define, generate, select, evaluate, and present explanations.

Intelligibility and Explainable Systems Research in HCI

The challenges of interaction with intelligent and autonomous systems have been discussed in the HCI community for decades. Pioneering work includes Suchman's Plans and Situated Actions [163], which critiqued AI's rigid concepts of plans and goals and pointed out its incompatibility with how people behave in the real world. Further discussions in the 90s include, for example, Shneiderman and Maes' CHI '97 panel on direct manipulation vs. interface agents [157].

In the late 90s and early 2000s, low-cost sensors and mobile devices drove research in context-aware computing forward [40, 42, 152]. Echoing earlier arguments from Suchman [163], researchers critiqued simplistic representations of context [53, 56]. It also became clear that people needed to be able to *understand* what was being sensed and which actions were being taken based on that information. Researchers argued for systems to provide accounts of their behaviour [13, 51, 52]. Notably, Bellotti and Edwards proposed that context-aware systems should be made *intelligible* by informing users about "what they know, how they know it, and what they are doing with that information" [13].

This fuelled further work in supporting intelligibility (or *scrutability* [8, 89]). Researchers proposed tailored interfaces that explained underlying context-aware rules [41], or provided textual [110, 111, 113, 178, 179] and visual explanations [177] for these rules. Other works explored how to design for implicit interaction with sensing-based systems [14, 75, 86, 175], and help people predict what will happen through feedforward [12, 176]. Another relevant stream of work in the intelligent user interfaces community explored how end-users make sense of and control machine-learned programs, working towards intelligible and debuggable machine-learned programs [100, 101]. The importance of understandability and predictability has also been recognized in interaction with autonomous vehicles [137, 153].

We see several other recent pockets of work in this area. Given the increasing amount of algorithmic decision making in society, HCI researchers have studied algorithmic transparency [26, 33], accountability [43] and fairness [108]. For instance, Diakopoulis looks at this from a computational journalism perspective [47] and curates a list of newsworthy algorithms used by the U.S. government [45]. Similarly, efforts have been made in the information visualization and visual analytics communities to visualize ML algorithms [90, 99, 109]. Unfortunately, the streams of research in explainable systems in the AI and ML communities and in the HCI community tend to be relatively isolated, which we demon-

strate in our analysis. Therefore, this work lays out the relevant domains involved in explanations and understanding and sets out an HCI research agenda for explainable systems.

Analyses on Trends in Research Topics

The traditional method to provide an overview of the state of the art and to assess a research topic is to perform a literature review. Examples in HCI are Jansen *et al*.'s research agenda for data physicalization [84], Chong *et al*.'s survey of device association [27], Pierce and Paulos' review of energy-related HCI [143], Froehlich *et al*.'s eco-feedback technology survey [59], Grosse-Puppendahl *et al*.'s capacitive sensing review [71] and Grossman's software learnability survey [73].

In this paper, we seek to examine trends across multiple academic domains, covering thousands of papers. Therefore, we use topic modelling, a semi-automated method to perform *literature analysis*. Bibliometric analysis has been used previously to characterize research areas. Most relevant to our work is Liu and colleagues' co-word analysis to analyze trends and links between different concepts for the Ubicomp [120] and CHI communities [121]. Co-word analysis [20] is an established literature analysis method that has been used to survey psychology [105], software engineering [31] and stem cell research [3]. It identifies clusters of keywords that often appear together in papers. For example, Liu and colleagues [120, 121], automatically extract keywords from papers to perform a keyword analysis. In this work, we analyze a much larger dataset (> 12,000 research papers) from multiple domains covering about 100 publication venues, instead of one conference. We performed Latent Dirichlet Allocation (LDA) based topic modeling [16] along with co-occurrence analysis to map out the research space. In summary, while our survey is HCI-centric, it covers a larger number of papers and a wider range of research areas than conventional literature reviews. We perform a data-driven literature analysis and carry out further analysis through visualization.

CITATION NETWORK

To build the citation network, we follow a semi-automated approach that combines the benefits of traditional literature reviews driven by the researchers' expertise with automated methods that help analyze a large body of literature. This also helps ameliorate concerns with keyword based search such as inconsistencies in keyword usage or missing entries [120, 121]. The manually curated set consists of 104 formative papers relating to explainable systems and explanations theory that were included based on the authors' expertise. Papers included based on keyword based search consists of 261 papers crawled from the Scopus [154] database by searching for the author or index keywords based on common variations in literature of the terms "intelligible", "interpretable", "transparency", "glass box", "black box", "scrutable", "counterfactuals" and "explainable". The resulting set of 365 papers was distilled down to 289 core papers by pruning those not relevant to our analysis. Specifically, we excluded papers about location transparency (networked resource abstraction), self-explanation (education), social transparency (translucence, awareness), context interpretation (inference from sensor data) and explanation based learning (machine learning), because they were not related to explaining or understanding systems and algorithms. Moreover, the core set of papers was revisited whenever expected topics based on the authors' judgement were missing, e.g., causal explanations, psychology of explanations. For, the core papers thus gathered, we crawled the citations from Google Scholar [66] for relevant metadata such as title, year of publication, citation count and publicly accessible PDF link. The final citation network consists of 289 core and 12412 citing papers.

Research Communities

Visualizing this citation network using Gephi [11] reveals different research communities and how they relate to one another. We identify 28 significant clusters (marked with **bold** names in Figure 1) after excluding smaller ones and summarize them in terms of 9 research communities. In the sections that follow, we ***bold-italicize*** the names of specific clusters represented in Figure 1 and describe smaller clusters within each community with *italicized* names. Two or more research clusters appear closer when many citing papers co-cite core papers from both clusters. Cluster's boundaries were determined visually from their separation in the visualization, but they were combined if the roots are similar (e.g., Interpretable Machine Learning and Classifier Explainers both have roots in ML). To verify if we omitted any significant clusters and to validate our labelling, we perform community extraction in Gephi based on modularity optimization [17] with resolution [104]. We set the resolution to 0.8, slightly favoring more albeit smaller clusters. We found that the resulting clusters are largely consistent with our labelling as shown in the figure included in the supplementary material. The network has a modularity score of 0.707 and 33 clusters. The 5 additional clusters are subsumed by our original labelling and stem from a finer resolution parameter.

Early Artificial Intelligence

Since the early promises of computing and artificial intelligence (AI), researchers have been concerned with understanding the systems' inner workings. In the late 1970s and 1980s, ***Expert Systems*** using knowledge bases [165] and ***Production Rules*** (e.g., [39]) were the first to add explanation capabilities, especially for medical decisions. Similarly, exploiting the intuitiveness of rules, ***Rule Extraction*** was developed to interpret artificial neural networks (ANN) and support vector machines (SVM), with work spanning from the 1990s to 2000s (e.g., [6, 139]). Later research focused on explaining ***Bayesian Belief Networks*** [102, 126] and ***Case-Base Reasoning*** systems [159], following the trends of research interest in these AI paradigms.

It is notable that while work on Rule Extraction remains active, it remains isolated [79]. In contrast, explanations of Expert Systems, Bayesian Networks, and Case-Base Reasoning influenced later research on explaining intelligent systems. However, we found that the recent burgeoning interest in interpretable ML is not well connected to this body of work.

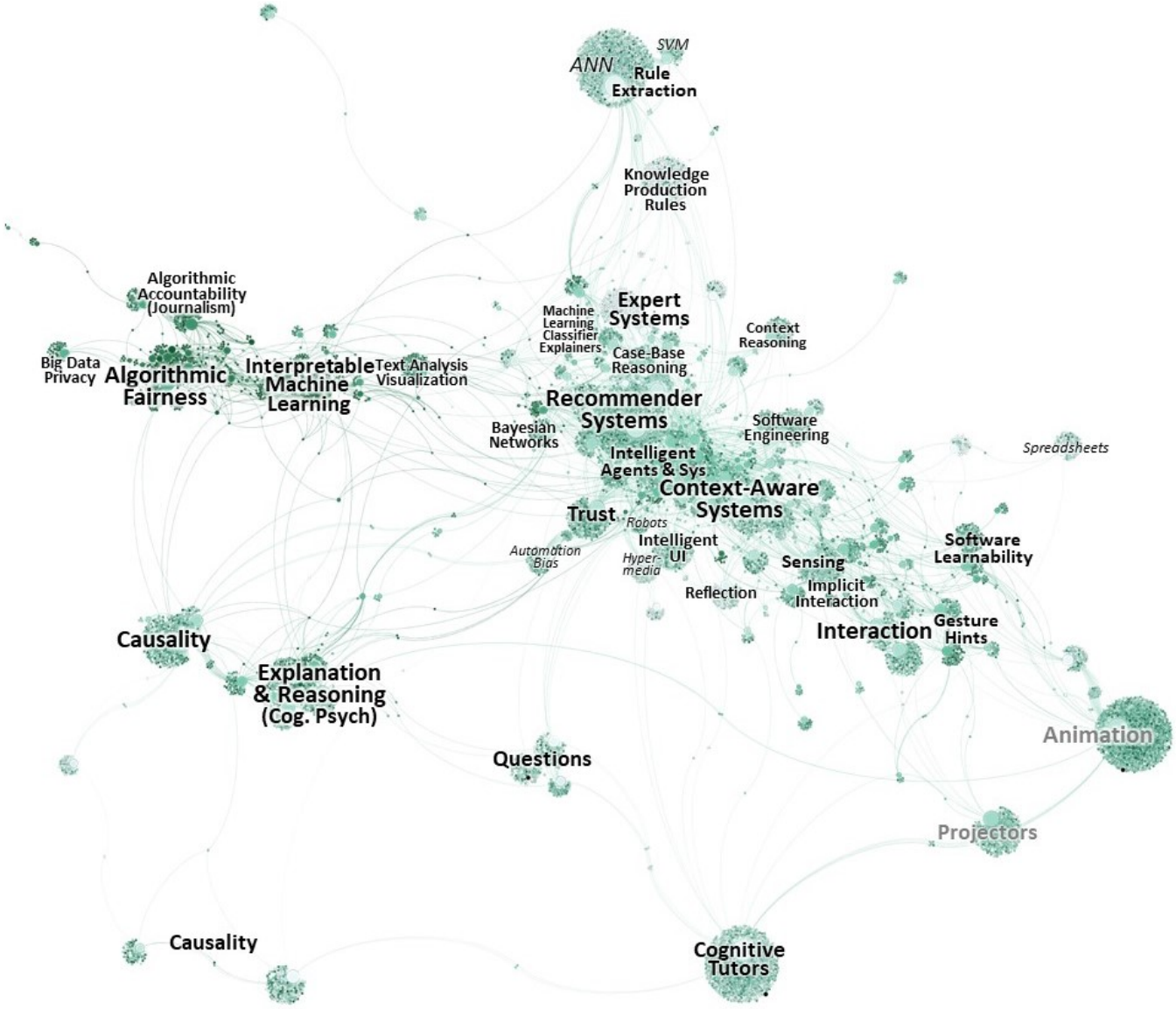

Figure 1. Citation network of 12,412 papers citing 289 core papers on explanations. This shows several communities of research domains, how they are closely related due to co-citations by citing papers and trends of research interest on explainable systems. Each node (circle) indicates a paper. Node size (of core papers) indicates the logarithm of the number of citing papers. Node color indicates the paper's age; darker green for more recently published papers. The nodes and edges are clustered by a force-directed layout method (ForceAtlas2 in Gephi [83]). Nodes are closer if more citing papers co-cite the same core papers. For example, recommender systems, intelligent agents and context-aware systems are closely related in terms of explainable systems, but the recent work on algorithmic fairness and interpretable classifiers are mostly developing independently. The size of the text roughly corresponds to the number of papers identified.

Intelligent Systems, Agents, and User Interfaces

The proliferation of personal computers and the Internet, from the mid-1990s drove the development of more applications for non-technical end-users and consumers. Given the broader user base, these systems would need to be understandable to non-technical and non-expert users to gain their ***Trust*** [55]. Hence, research on explainable systems focused on ***Intelligent Systems*** [68] and ***Intelligent Agents*** [62, 80] to help with decision making or task management.

Some research focused on consumer-oriented applications, such as ***Recommender Systems*** [33, 81] and *Adaptive User Interfaces* of applications and for ***Software Engineering*** [95, 96, 135]. With the increasing use of machine learning, ***Intelligent User Interfaces*** began to focus on providing explanations to *debug learned models* [100, 101, 162]. Explaining the increased level of complexity is also addressed in research on ambient intelligent systems.

Ambient Intelligence: Sensing and Context-Awareness

The miniaturization of electronics drove the development of ***Sensors*** and *Pervasive Systems* using ***Context-Awareness*** to recognize user activity and intent. These systems bring additional challenges for user trust and understanding [13, 169]

due to the ***Implicit Interaction*** nature of ubiquitous sensing, and increasingly complex reasoning mechanisms.

The heterogeneity of ubiquitous computing and increasing complexity of systems poses several challenges for interaction. Researchers have proposed conceptual frameworks to help address these challenges, such as seamful design [23], design questions to understand and control sensing systems drawn from the social sciences [14], and computational ***Reflection*** and abstraction [51].

To support these goals, several software toolkits have been proposed to aid the development of explainable ambient intelligence: PersonisAD for scrutable personalization [8], PersonisLF for long-term user models with forgetting [10], Encators to explain context reasoning [41], and the Intelligibility Toolkit to explain question types [113].

Research explored the need for [111] and provision of explanations in physical space, such as smart environments [177, 178, 179], the office [26], smart homes [18, 30, 131, 183] and on the move [90, 116]. The fallibility of sensing and inference systems presents challenges to understanding, that can be addressed by visualizing uncertainty [7, 90, 115, 151].

Interaction Design and Learnability

Branching off from the work by Bellotti and colleagues to make sense of sensing systems [14], we find research clusters around explainable systems that can be categorized as ***Interaction Design***. Many novel interaction techniques enabled by sensor-driven systems lacked discoverability (e.g., gestural interaction on tabletops or cross-device interaction) and feedback (e.g., proactive displays). Additionally, researchers were building increasingly sophisticated ubicomp spaces, consisting of multiple communicating distributed components such as sensors and devices. Researchers explored solutions to these issues, e.g., by presenting what devices and services are available through *spontaneous discovery* [60], with the concept of meta-user interfaces to configure ambient spaces [32], by exploring ways to facilitate cross-device interaction with gradual engagement [128] and with a conceptual model and interaction techniques for systems that increasingly reside and act in the periphery (e.g., [86, 87]). This body of work also includes research on how feedforward can help people understand and predict what is going to happen, initially for tangible interaction from a product design angle [48, 49, 180], and later for gestural interaction (e.g., [12, 58]) and from a broader HCI perspective [176].

Another research cluster with a rich history since the 70s is ***Software Learnability*** as a key component of usability [72, 73]. This body of work investigates solutions to aid learning how to use complex software applications (e.g., AutoCAD), with techniques such as in-context videos or demonstrations [72]. Note that the Software Learnability cluster and the clusters on the effectiveness of ***Animation*** (e.g., [140, 170]) and ***Projectors*** [145] stem from the manually curated core papers. We hypothesized that these could be relevant strategies to consider in XAI research. However, our analysis and the visualization in Figure 1 objectively show that these clusters (labelled in grey) are relatively distant to the central clusters, highlighting our method's robustness against work that turns out to be rather tangential. Nevertheless, this may also suggest opportunities for cross-pollination of ideas between these clusters and XAI research.

Interpretable ML and Classifier Explainers

Machine learning (ML) has had a long history spanning decades and there are many subareas in ML, but research on explaining ML has primarily focused on classification. Early work on machine learning classification from the mid-2000s explored explanations using Nomograms [133], visual additive classifiers [144], and explanations of instances [147].

The recent strong interest in AI can be attributed to *Deep Learning* [64] which has made dramatic strides in areas such as computer vision and natural language. However, just as Artificial Neural Networks were previously criticized as being uninterpretable, so is Deep Learning. This has driven a lot of interest to explain various types of deep neural networks (DNNs). Some methods attempt to explain DNNs, such as attention-based models for visual Question-Answering (QA) [36] and Grad-CAM [155] for producing visual explanations of convolutional neural network (CNN) based models by highlighting regions relevant to target tasks.

While some methods seek to develop "glass box" models which are intrinsically interpretable (e.g., generalized additive models [22], Bayesian rule lists [110, 160]), other methods work as a proxy explainer over "black box" models (e.g., sparse linear model [146]). Kim identifies three main categories of interpretable models [91]: sparse linear classifiers [146, 160], discretization methods such as decision trees and association rule lists [103, 110], and instance- or case-based models [92, 94].

Algorithmic Fairness, Accountability, Transparency, Policy, and Journalism

The prevalence of algorithmic decision-making systems in society has spurred the need for ***Fairness*** in systems. Hajian called for data mining systems which are discrimination-conscious by-design [76]; such as fair systems for loan risk prediction [141], and bail flight risk of violent crime suspects [29], as well as methods to detect bias in black box models, such as Quantitative Influence (QI) [38] and methods that guarantee fairness such as contextual bandits from game theory [85]. Fairness is also important when algorithms are used in social settings to support group decision making [109], or the allocation of distributed work [107]. Lee *et al*. observed that notions of fairness vary for different stakeholders and do not always correspond to mathematical interpretations of algorithmic fairness [108].

Other than being fair, algorithms also need to be ***Accountable*** and explain their decisions. This will soon become law in some countries, e.g., the EU "right to explanation" [65]. Shneiderman argues for an independent process, including planning oversight, continuous review and retrospective

analyses to hold algorithms accountable [156]. Other authors have also argued for the need to define *policy* [54]. ***Privacy*** issues due to the advent of big data analytics [147,166] and the contention between transparency and privacy is also an active area of investigation [38, 111].

Journalism can also play a role to hold algorithms accountable by reporting issues of bias and algorithmic power to the broader public [43]. Moreover, as news media becomes increasingly automated, the use of algorithmic curation and automated writing leads to a greater need for transparency [46].

Causality

While machine learning typically models correlations, ***Causality*** is concerned with establishing the cause-effect relationships. Causal discovery utilizes the theory of Bayesian networks to discover causal structures in data while causal explanations provide a reasoning for why the events occurred. Pearl's pioneering work on structural causal models [77, 78] provides a unifying computational framework for causal reasoning and explanations based on counterfactuals. A counterfactual is a conditional statement contrary to a fact. For example, if event A would not have happened then event B would not have happened.

Psychological Theories of Explanations

Researchers from ***Psychology*** have developed a rich theory of different types of explanations (functional, mechanistic, causal [123, 124]) and explored their role in learning, reasoning, categorization of knowledge [122] and scientific understanding [168].

Education and Cognitive Tutors

In the field of education, a lot of attention has been paid to how students learn by asking ***Questions*** and the question generation mechanisms [67]. Closely related to this, the field of ***Cognitive Tutoring*** from the 90's based on the ACT [4] cognitive model has explored the role of explanations in how students learn in classrooms and applied it to the development of computer based tutors [5].

TOPIC NETWORK

With the citation network, we can identify ***which*** communities are connected or isolated and identify the trends over time. Next, we want to understand ***what*** the key topics in each community are, ***how*** closely coupled communities relate to each other and ***why*** citing papers cited the core papers. Therefore, we partition the network into four subnetworks containing communities that are close together and perform topic modeling on the abstracts of the core papers, the abstracts of the citing papers, and the paragraphs of the citing papers where the core papers are cited. The intuition behind this is that topics in the abstract are about the paper in general whereas the citing paragraph would contain more contextual information that relates the core paper to the cited paper.

Topic Modeling

With the 289 core and 12,412 citing papers identified, we attempted to download the openly accessible PDF where available. The downloaded PDFs were then processed using Grobid [125] to extract the abstract (both core and citing papers) and the citing paragraph (from the citing paper). The extracted text from the citing abstracts and citing paragraphs were further processed by removing non-alphanumeric characters, stemming plural forms to make them singular and stop word removal. After preprocessing, we had 289 core abstracts, 6597 cite abstracts and 10,891 cite paragraphs. Compared to the citation network we have approximately 47% fewer citing papers since a) not all PDFs are publicly available to be retrieved by the automated crawling process and b) we limit our analysis to English language texts.

To focus the discussion on central research clusters, consolidate higher-level themes, and understand how proximate clusters relate to each other, we analyze four subnetworks from the citation network as shown in Figure 1. Starting from the bottom left quadrant and moving in a clockwise direction, we notice causality and psychology of explanations. The upper left quadrant consists of algorithmic fairness, transparency and interpretable machine learning. The central portion of the upper right quadrant consists of the densest of the subnetworks, relating to intelligent and ambient systems. Finally, the central portion of the bottom right quadrant consists of interactions and software learnability and is tightly connected to ambient and intelligent systems. For each of the subnetworks, we performed the following steps:

1) Iterative Open Coding of Topics in Core Abstracts: The authors of this paper reviewed the core abstracts to label the core papers with the identified topics. Our labels were then combined with the author-provided keywords of the core papers. This final set of keywords were iteratively refined until a concise set of topic keywords emerged.

2) Topic Modelling of Citing Abstracts: We performed LDA based topic modelling utilizing Gibbs sampling [70] on the preprocessed text of citing abstracts. LDA is a generative probabilistic model. Given a bag of words representation of documents, LDA treats each document as a distribution over topics and each topic as a distribution over words. LDA requires the number of topics as one of the inputs to learn the topic distributions. While there are metrics available to help with selection of number of topics, they are far from perfect beyond standard benchmark datasets. To determine the number of topics, we use the R package ldatuning [136], which implements various metrics, as a starting point and iteratively perform topic modelling until the number of topics provides a good tradeoff between too general and overtly specific. LDA generates unlabeled topics. To label them, we generate the top 30 words for each topic and manually labeled them. Multiple coauthors inspected the labels to converge on the final agreed labels.

3) Topic Modelling of Citing Paragraphs: This is the same as step 2, but for preprocessed texts from citing paragraphs.

4) Topic Network Generation: After the topic modelling and labelling is complete, we have a set of topics for each of

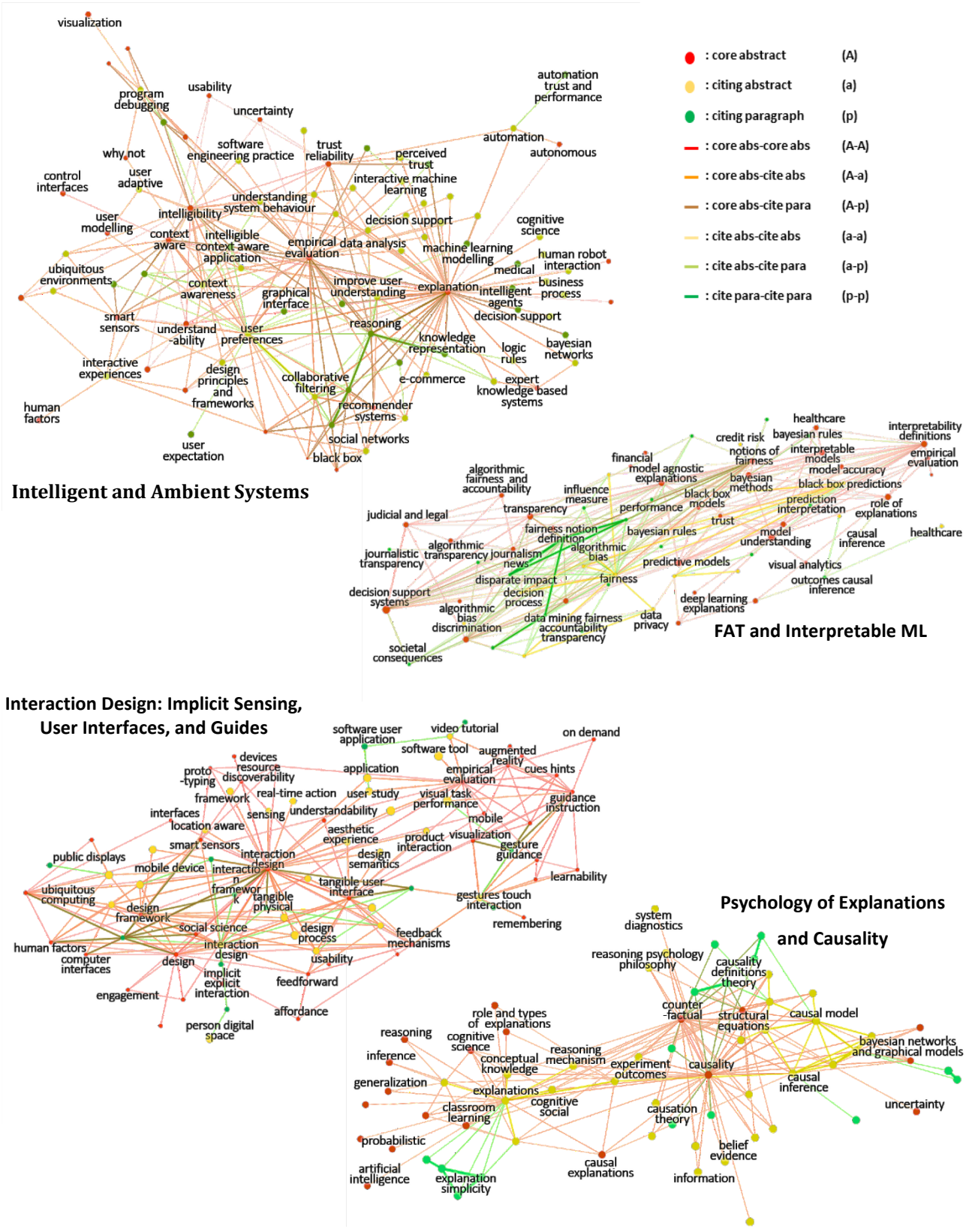

Figure 2. Topic networks of four key subnetworks of core and citing papers. Topics derived from LDA modelling, which were manually labeled by the authors. Nodes indicate topics of different types: red Core Abstract (A) topics, yellow Cite Abstract (a) topics, green Cite Paragraph (p) topics. Edges indicate inter-document (aA, pA, pa) connections or intra-document (aa, AA, pp) co-occurrences of topics. A document is a body of text either core abstract, citing abstract or citing paragraph. Thicker edges indicate more co-occurrences, which also causes nodes to be closer together. Edges have been filtered to balance between too many nodes/edges which produce a non-interpretable "hairball", and sufficient nodes/edges to draw insights.

the three sources of texts: core abstracts, citing abstracts and citing paragraphs. We then use the respective trained LDA models for each text and predict up to five most likely topics that occur with a probability greater than a uniform distribution on the number of topics. To understand the relationship between these topics, we create a topic network by constructing an undirected graph as follows:

- **Nodes:** Each unique topic from each of the sources of texts is treated as a node. We have three types of nodes, core abstract topics colored red, cite abstract topics colored yellow and cite paragraph topics colored green.
- **Edges:** We have a total of six different types of edges weighted by the frequency of co-occurrence
  a. **Co-occurrence edges:** There are three types of co-occurrence edges: core abstract to core abstract colored red, cite abstract to cite abstract colored yellow and cite paragraph to cite paragraph colored green. This encapsulates which topics within a set of texts co-occur.
  b. **Network edges:** There are three types of network edges. Cite abstract to core abstract, colored orange, which indicate the relationship between topics in the core and citing papers. Cite paragraph to core abstract colored brown and cite paragraph to cite abstract colored light green, which provide context on why a citing paper cites a core paper.

Topic Network Analysis: The resulting topic network was then visualized using the ForceAtlas2 algorithm [83] in Gephi [11]. The graphs generated are dense and contain rich contextual information. To ease interpretability, we filter the edges based on weights and present the most dominant patterns in the various subnetworks as shown in Figure 2. The size of the nodes is proportional to the occurrence of topics. A bigger node indicates a more prevalent topic.

Topic Groups: Sub-Networks

We describe some insights from the topic modeling, and discuss how research topics relate within the subnetwork and how the subnetworks relate to one another.

Intelligent and Ambient (I&A) Systems

This is the biggest, most central, and mature subnetwork spanning many research areas of the Intelligent Systems and Ambient Intelligence communities. The proximity of these clusters to each other indicates the cross-pollination of ideas between the two research areas. For example, the use of Why questions for explanations originated in research in user-centered software engineering [95, 96] but is commonly used for explanations in both communities [101, 113, 117, 135, 179]. Their objectives were also similar – to improve the end-user trust and usability of these systems, whether as a desktop or pervasive interface. The subnetwork has its roots from Expert Systems to Recommender Systems, Context-Aware Systems, and Intelligent User Interfaces. Some recent developments include generating verbalization explanations for human-robot interaction [149] and interactive visualizations for iML models for data scientists [99].

From the topic network (Figure 2) we notice that research in this space is concerned with building compelling smart applications with a strong emphasis on user validation and empirical testing. Domains include recommenders [33], decision support [62], medical [19, 99], business processes [88, 98], e-commerce [164, 186], social networks [63, 161], smart sensors [112, 113], smart homes [2, 30, 183], games [134, 182], and public displays [179]. The user-centered objectives drive the need for control interfaces, user adaptation, and preference modeling. Specific concerns that require explanations include usability, trust, reliability, and understanding system behavior.

Explainable AI: Fair, Accountable, and Transparent (FAT) algorithms and Interpretable Machine Learning (iML)

For this topic subnetwork, we see that research on FAT is mainly driven by *societal issues* concerning black box systems, while iML is mainly interested in developing *methods* for interpretability in ML. Key issues for algorithmic fairness and accountability are: bias, discrimination in big data and algorithms, ethics, legal and policy implications, disparate impact, data privacy, and journalistic transparency. Key methods for interpretable machine learning are: Bayesian rules, explanations for deep learning, visual analytics and causal inference.

While I&A Systems target many consumer applications, the applications considered for iML are usually from more critical domains, such as medical [22, 99, 160] or finance, and applications for FAT are more wide-reaching, such as judiciary, policy and journalism. Unlike with I&A Systems, where there is a strong emphasis on validation with real users and scenarios, validation for FAT and iML appear to be primarily performed on commonly available datasets.

Theories of Explanations: Causality & Cognitive Psychology

In stark contrast to the other research communities, Causality and Cognitive Psychology both focus on the theory of explanations. While the work on causality and Bayesian networks spearheaded by Halpern and Pearl [77, 78] approaches the problem from a computational perspective, the work by Lombrozo [122, 123, 124] explores the cognitive aspects of explanations. Although seemingly disjoint, these communities two are strongly connected by counterfactual reasoning and causal explanations.

Interactivity and Learnability

We found a broad emphasis on interaction design (central topic in this subnetwork), design principles, interaction styles, and software learnability in this subnetwork. We can see three main trends from this topic network: software learnability (*top right*), design principles (*bottom middle*), and interface technology and interaction styles (*left*).

The topics centered around interface technology and interaction styles relate to exploring the design space of ubiquitous computing beyond the desktop, including mobile devices, tabletop surfaces, public displays, gestures, tangible, and physical interfaces. Sensor-driven interaction is an important

aspect of this (top left). As part of research into these new interface technologies, we see supporting research related to interaction design, such as design frameworks, implicit and explicit interaction (e.g., [86]).

On the right, we see topics that mostly relate to software learnability, such as cues, hints, tutorials, on-demand, guidance and visualizations. Much of this is motivated by novel interface technology with challenges in usage and understanding what is possible, i.e., gestural interaction and touch interfaces. There is also a strong link to evaluation, suggesting that these tutorials, guides and visualizations tend to be empirically evaluated through user studies.

Finally, we observe topics that deal with the design process and core design principles such as feedback (mechanisms), affordances, and feedforward [176]. Particularly feedback is a core concepts within HCI [138], that we can see as part of our vocabulary to discuss challenges in the use of interactive technology (i.e., the need for appropriate feedback [138]). This suggests that these principles are central HCI concepts in terms of challenges of understanding (e.g., see the framing in [14]), with feedforward [48, 176] becoming more relevant.

In terms of vocabulary, the algorithmic fairness and accountability community and this subnetwork are quite different. Even though there are overlaps between venues between these subnetworks (e.g., CSCW work on algorithmic fairness in the sharing economy [108]), the fairness and accountability community are shown to explore problems more from a big picture perspective and focus more on problems with a potential large societal impact. In contrast, the interaction design subnetwork appears to focus more on understanding and learnability of novel or complicated interfaces. This illustrates an opportunity to bridge this gap and cross-pollinate ideas from both communities.

Finally, as a research domain, HCI research is interdisciplinary and regularly draws insights and inspiration from other domains. However, rather curiously, Figure 1 shows that the research clusters on interaction design are quite distant from and appear to be less informed by relevant theoretical domains in computer science and cognitive psychology. It could imply that there are knowledge gaps, or could point to the lack of practical applicability of relevant theoretical concepts from these other domains.

DISCUSSION AND IMPLICATIONS

Based on our analysis of the citation and topic networks, we articulate trends, trajectories and research opportunities for HCI in explainable systems. We articulate insights based on: (i) reading and analyzing sample papers from different communities; (ii) closeness of communities indicated by separation in Figure 1; (iii) separation over time indicated by color intensity of nodes in Figure 1; (iv) topic words discovered in sub communities in Figure 2; and, (v) the authors expertise.

Trend: Production Rules to Machine Learning

Interestingly, we found that the earlier methods to explain machine learning models (Classifier Explainers) were more strongly related to the Expert Systems and Bayesian Networks clusters, rather than the new cluster on Interpretable ML. This suggests that the new approaches could (i) be using different techniques and the older methods are obsolete, (ii) be targeting very different problems, or (iii) have neglected the past. Furthermore, while rule extraction is an old research area, and quite separated from the recent developments in iML and FAT, there have been new papers within the past few years, such as methods for rule extraction from deep neural networks [187, 75]. Therefore, we should reflect on the past research topics to rediscover methods that could be suitable for current use cases.

Trend: Individual Trust to Systematic Accountability

While research on explainable Intelligent and Ambient Systems (I&A) and Interpretable Machine Learning (iML) have focused on the need to explain to single end-users to gain their *individual trust* [13, 55, 68], research on fair and accountable, and transparent (FAT) algorithms aims to address *macroscopic societal accountability* [44, 156]. There is a shift in perceived demand for intelligibility from the individual's need for understanding, to their need for institutional trust. This would require understanding requirements arising from social contexts other than just form usability or human cognitive psychology. Therefore, it is important to draw insights from social science too [132].

Trajectory: Road to Rigorous and Usable Intelligibility

Explainable AI (including FAT and Interpretable Machine Learning) focuses on the mathematics to transform a complex or "black box" model into a simpler one, or create mathematically interpretable models. While these are significant contributions, they tend to neglect the human side of the explanations and whether they are usable and practical in real-world situations [50, 119, 132]. As shown in Figure 2, the research does not appear to be strongly informed by Cognitive Psychology in terms of how humans can interpret the explanations, and does not deploy or evaluate the explanations in interactive applications with real users. This presents an opportunity for HCI research to bridge this gap.

Research in HCI on explainable interfaces has demonstrated the value of explanations to users with classical machine learning models (e.g., linear SVM, naïve Bayes) [101, 162]. These are good starting points to understanding how users use explanations. However, real world applications and datasets have challenges which require more sophisticated models (e.g., missing data, high dimensionality, unstructured data). Furthermore, sometimes the reason the system does not work may lie outside the system, which may require the provision of additional information that is not related to the internal mechanics of the system.

To **improve rigor**, as we test the effectiveness of explanation interfaces, we should use real-world data with functional complex models that deal with intricacies. We can draw from the diverse models and algorithms being developed by the FAT and iML research clusters. On the other hand, causal

explanations [77, 78] could provide a strong theoretical framework to reason about and generate explanations.

To **improve usability** of intelligible or explainable interfaces, we can draw from the rich body of research in HCI on Interaction Design (e.g., [86]) and Software Learnability (e.g., [73]). Furthermore, we can draw from theoretical work on the Cognitive Psychology of explanations [124] to develop explanations that are easier to interpret. Finally, HCI researchers can perform empirical studies to evaluate the efficacy of the novel explanation interfaces in various settings (e.g., [114, 116]).

Trajectory: Interaction and Interfaces for Intelligibility

An important area for future work is exploring **interactive explanations** and **interfaces for intelligibility**. Most explanations resulting from research in the XAI community are *static* (e.g., [146]) and assume that there is a single message to convey through the explanation. An alternative approach would be to allow users to explore the system's behaviour freely through interactive explanations.

As shown in the visual analytics [28] and information visualization communities [21], interaction can be a powerful means to enable people to iteratively explore and gather insight from large amounts of complex information. This suggests that allowing people to interactively explore explanations for algorithmic decision-making is a promising direction. We already see some examples of research in this direction, such as studies on understanding how data scientists work [97], tools to support data scientists in using and understanding machine-learning algorithms [99, 130, 142, 141], visualization techniques for text analysis [34, 35], tools to support interactive data analysis [24, 25, 37] and interactive machine learning [82]. However, more work is needed to tailor these interfaces to different audiences and exploit interactivity (e.g., while Kay *et al*.'s transit visualizations were aimed at non-experts, they were non-interactive [90]).

To go beyond static explanations, researchers can draw from existing work on intelligibility for context-aware systems, including design space explorations (e.g., [173, 174]), conceptual models for implicit interaction [86] and intelligible interfaces for different scenarios and using various modalities (e.g., [111, 175, 177]).

Finally, an interesting direction is work on effectively interacting with AI augmentation tools. A recent example is a technique that provides interactive visualizations of a gesture recognizer's recognition rate to let people design their own gestures that are also easy to recognize [127].

LIMITATIONS

While we have made every effort to ensure a broad coverage of papers relevant to explainable systems, we had yet to analyze other research domains which cover explanations, such as theories from philosophy and social sciences, Bayesian and constraint-based cognitive tutors, relevance feedback and personalization in information retrieval.

It is possible that the process of manual curation of the initial set of core explanation papers may have introduced a bias. This could be improved by iterative citation tracing, where citing papers identified as central to explanation systems could be added to core explanation papers; and by backward reference searching (chain searching) to discover root explanation papers of core explanation papers. Such an iterative process would provide further evidence for gaps or collaborations between the various communities.

Our topic modeling led to many nodes that we filtered out to improve readability and interpretability. However, this filtered out interesting but less frequent topics, such as the need for intelligibility in energy-efficient smart homes to explain smart thermostats [183] or smart laundry agents [30]. Nevertheless, the semi-automated analysis methods allowed us to capture important trends and relationships across many papers and spanning many domains.

We assume that the citing paragraph is relevant to the core paper that it cites. While this assumption could be questioned (e.g., Marshall reported on shallow citation practices in CHI papers [129]) the bibliometric analysis method that we employed, including the use of citation network visualizations, is well-established. Given the scale of our citation network spanning multiple domains, we also expect that our results will be less sensitive to noise.

CONCLUSION

Recent advances in machine learning and artificial intelligence have far-reaching impacts on society at large. While researchers in the ML and AI communities are working on making their algorithms explainable, their focus is not on usable, practical and effective transparency that works for and benefits people. Given HCI's core interest in technology that empowers people, this is a gap that we as a community can help to address, to ensure that these new and powerful technologies are designed with intelligibility from the ground up. From a literature analysis of 12,412 papers citing 289 core papers on explainable systems, we mapped the research space from diverse domains related to explainable systems. We revealed fading vs. burgeoning trends and connected vs. isolated domains, and from this, extracted several implications, future directions and opportunities for HCI researchers. While this is only a first step, we argue that true progress towards explainable systems can only be made through interdisciplinary collaborations, where expertise from different fields (e.g., machine learning, cognitive psychology, human-computer interaction) is combined and concepts and techniques are further developed from multiple perspectives to move research forward.

ACKNOWLEDGEMENTS

We thank Shuqi Wang for her assistance in coding some of the papers. This work was carried out in part at the SeSaMe Centre, supported by Singapore NRF under the IRC@SG Funding Initiative and administered by IDMPO.

REFERENCES

1. Aha, D. W., Darrell, T., Pazzani, M., Reid, D., Sammut, C., & Stone, P. (2017). Workshop on Explainable AI (XAI). *IJCAI 2017*.
2. Alan, A. T., Costanza, E., Ramchurn, S. D., Fischer, J., Rodden, T., & Jennings, N. R. (2016). Tariff agent: interacting with a future smart energy system at home. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *23*(4), 25.
3. An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. Scientometrics, 88(1), 133-144.
4. Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.
5. Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences*, *4*(2), 167-207.
6. Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-based systems*, *8*(6), 373-389.
7. Antifakos, S., Kern, N., Schiele, B., & Schwaninger, A. (2005, September). Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services* (pp. 9-14). ACM.
8. Assad, M., Carmichael, D. J., Kay, J., & Kummerfeld, B. (2007, May). PersonisAD: Distributed, active, scrutable model framework for context-aware services. In *International Conference on Pervasive Computing* (pp. 55-72). Springer, Berlin, Heidelberg.
9. Barocas, S., Friedler, S., Hardt, M., Kroll, J., Venkatasubramanian, S., & Wallach, H. The FAT-ML workshop series on Fairness, Accountability, and Transparency in Machine Learning. Accessed on 2017-09-09. http://www.fatml.org/
10. Barua, D., Kay, J., Kummerfeld, B., & Paris, C. (2011, November). Theoretical foundations for user-controlled forgetting in scrutable long term user models. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference* (pp. 40-49). ACM.
11. Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, *8*, 361-362.
12. Bau, O., & Mackay, W. E. (2008, October). OctoPocus: a dynamic guide for learning gesture-based command sets. In *Proceedings of the 21st annual ACM symposium on User interface software and technology* (pp. 37-46). ACM.
13. Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: human considerations in context-aware systems. *Human–Computer Interaction*, *16*(2-4), 193-212.
14. Bellotti, V., Back, M., Edwards, W. K., Grinter, R. E., Henderson, A., & Lopes, C. (2002, April). Making sense of sensing systems: five questions for designers and researchers. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 415-422). ACM.
15. Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)* (p. 8).
16. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
17. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.
18. Bourgeois, J., Van Der Linden, J., Kortuem, G., Price, B. A., & Rimmer, C. (2014, September). Conversations with my washing machine: an in-the-wild study of demand shifting with self-generated energy. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 459-470). ACM.
19. Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of 2015 International Conference on Healthcare Informatics (ICHI)*, (pp. 160-169). IEEE. (Healthcare in IUI).
20. Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, 22(2), 191-235.
21. Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). Readings in information visualization: using vision to think. Morgan Kaufmann.
22. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730). ACM.
23. Chalmers, M., & MacColl, I. (2003, October). Seamful and seamless design in ubiquitous computing. In *Workshop at the crossroads: The interaction of HCI and systems issues in UbiComp* (Vol. 8).
24. Chau, D. H., Kittur, A., Hong, J. I., & Faloutsos, C. (2011, May). Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI Conference*

on Human Factors in Computing Systems (pp. 167-176). ACM.

25. Chau, P., Vreeken, J., van Leeuwen, M., Shahaf, D., & Faloutsos, C. (2016). Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics. In ACM SIGKDD 2016 Full-day Workshop on Interactive Data Exploration and Analytics. IDEA'16.
26. Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C., & Villar, N. (2005). Exploring issues of user model transparency and proactive behaviour in an office environment control system. *User Modeling and User-Adapted Interaction*, *15*(3), 235-273.
27. Chong, M. K., Mayrhofer, R., & Gellersen, H. (2014). A survey of user interaction for spontaneous device association. *ACM Computing Surveys (CSUR)*, *47*(1), 8.
28. Cook, K. A., & Thomas, J. J. (2005). Illuminating the path: The research and development agenda for visual analytics.
29. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. arXiv preprint arXiv:1701.08230.
30. Costanza, E., Fischer, J. E., Colley, J. A., Rodden, T., Ramchurn, S. D., & Jennings, N. R. (2014, April). Doing the laundry with agents: a field trial of a future smart energy system in the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 813-822). ACM.
31. Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: a study in co-word analysis. J. Am. Soc. Inf. Sci. 49, 13 (November 1998), 1206-1223.
32. Coutaz, J. (2006, October). Meta-user interfaces for ambient spaces. In International Workshop on Task Models and Diagrams for User Interface Design (pp. 1-15). Springer, Berlin, Heidelberg.
33. Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., ... & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455.
34. Chuang, J., Manning, C. D., & Heer, J. (2012, May). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces* (pp. 74-77). ACM.
35. Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012, May). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443-452). ACM.
36. Das, A., Agrawal, H., Zitnick, C. L., Parikh, D., & Batra, D. (2016). Human attention in visual question answering: Do humans and deep networks look at the same regions?. arXiv preprint arXiv:1606.03556.
37. Das, S., McCarter, A., Minieri, J., Damaraju, N., Padmanabhan, S., & Chau, D. H. P. ISPARK: Interactive Visual Analytics for Fire Incidents and Station Placement. In *ACM SIGKDD Workshop on Interactive Data Exploration and Analytics* (p. 29).
38. Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 598-617). IEEE.
39. Davis, R., Buchanan, B., & Shortliffe, E. (1977). Production rules as a representation for a knowledge-based consultation program. *Artificial intelligence*, *8*(1), 15-45.
40. Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing*, *5*(1), 4-7.
41. Dey, A. K., & Newberger, A. (2009, April). Support for context-aware intelligibility and control. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 859-868). ACM.
42. Dey, A. K., Abowd, G. D., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, *16*(2), 97-166.
43. Diakopoulos, N. (2014). Algorithmic Accountability Reporting: On the Investigation of Black Boxes. *Columbia University Academic Commons.*
44. Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (January 2016), 56-62.
45. Diakopoulis, N. (2018). *Algorithm Tips – Find tips for stories on algorithms*. *Algorithmtips.org*. Retrieved 8 January 2018, from http://algorithmtips.org/
46. Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, 5(7), 809-828.
47. Diakopoulos, N., & Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*,
48. Djajadiningrat, T., Overbeeke, K., & Wensveen, S. (2002, June). But how, Donald, tell us how?: on the creation of meaning in interaction design through feedforward and inherent feedback. In Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques (pp. 285-291). ACM.
49. Djajadiningrat, T., Wensveen, S., Frens, J., & Overbeeke, K. (2004). Tangible products: redressing the

balance between appearance and action. *Personal and Ubiquitous Computing*, *8*(5), 294-309.

50. Doshi-Velez, F., & Kim, B. (2017). A Roadmap for a Rigorous Science of Interpretability. arXiv preprint arXiv:1702.08608.
51. Dourish, P. (1995). Developing a reflective model of collaborative systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *2*(1), 40-63.
52. Dourish, P. (1997). Accounting for system behaviour: Representation, reflection and resourceful action. *Computers and Design in Context, MIT Press, Cambridge, MA, USA*, 145-170.
53. Dourish, P. (2004). What we talk about when we talk about context. *Personal and ubiquitous computing*, *8*(1), 19-30.
54. Dwork, C., & Mulligan, D. K. (2013). It's not privacy, and it's not fair. Stan. L. Rev. Online, 66, 35.
55. Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697-718.
56. Erickson, T. (2002). Some problems with the notion of context-aware computing. *Communications of the ACM*, *45*(2), 102-104.
57. European Union General Data Protection regulation (GDPR). 2016. Accessed on 2017-09-09. http://www.eugdpr.org/
58. Freeman, D., Benko, H., Morris, M. R., & Wigdor, D. (2009, November). ShadowGuides: Visualizations for in-situ learning of multi-touch and whole-hand gestures. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces* (pp. 165-172). ACM.
59. Froehlich, J., Findlater, L., & Landay, J. 2010. The design of eco-feedback technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10). ACM, New York, NY, USA, 1999-2008.
60. Gellersen, H., Fischer, C., Guinard, D., Gostner, R., Kortuem, G., Kray, C., ... & Streng, S. (2009). Supporting device discovery and spontaneous interaction with spatial references. *Personal and Ubiquitous Computing*, *13*(4), 255-264.
61. Gillies, M., Fiebrink, R., Tanaka, A., Garcia, J., Bevilacqua, F., Heloir, A., ... & d'Alessandro, N. (2016, May). Human-Centred Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '16). ACM, New York, NY, USA, 3558-3565.
62. Glass, A., McGuinness, D. L., & Wolverton, M. (2008, January). Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces* (pp. 227-236). ACM.
63. Golbeck, J., & Hendler, J. (2006, January). Filmtrust: Movie recommendations using trust in web-based social networks. In *Proceedings of the IEEE Consumer communications and networking conference* (Vol. 96, No. 1, pp. 282-286).
64. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Book in preparation for MIT Press. http://www.deeplearningbook.org.
65. Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.
66. *Google Scholar*. (2018). *Google Scholar*. Retrieved 7 January 2018, from http://scholar.google.com/
67. Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American educational research journal*, *31*(1), 104-137.
68. Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, 497-530.
69. Greis, M., Hullman, J., Correll, M., Kay, M., & Shaer, O. (2017, May). Designing for Uncertainty in HCI: When Does Uncertainty Help?. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17). ACM, New York, NY, USA, 593-600.
70. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101* (suppl 1), 5228-5235.
71. Grosse-Puppendahl, T., Holz, C., Cohn, G., Wimmer, R., Bechtold, O., Hodges, S., ... & Smith, J. R. (2017, May). Finding Common Ground: A Survey of Capacitive Sensing in Human-Computer Interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3293-3315). ACM.
72. Grossman, T., & Fitzmaurice, G. (2010, April). ToolClips: an investigation of contextual video assistance for functionality understanding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1515-1524). ACM.
73. Grossman, T., Fitzmaurice, G., & Attar, R. (2009, April). A survey of software learnability: metrics, methodologies and guidelines. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 649-658). ACM.
74. Gunning, D. (2017). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA).* Accessed on 2017-09-09. http://www.darpa.mil/program/explainable-artificial-intelligence

75. Hailesilassie, T. (2016). Rule extraction algorithm for deep neural networks: A review. *arXiv preprint arXiv:1610.05267*.

76. Hajian, S., Bonchi, F., & Castillo, C. (2016, August). Algorithmic bias: from discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2125-2126). ACM.

77. Halpern, J. Y., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science*, *56*(4), 843-887.

78. Halpern, J. Y., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British journal for the philosophy of science*, *56*(4), 889-911.

79. Hayashi, Y., Setiono, R., & Azcarraga, A. (2016). Neural network training and rule extraction with augmented discretized input. *Neurocomputing*, *207*, 610-622.

80. Haynes, S. R., Cohen, M. A., & Ritter, F. E. (2009). Designs for explaining intelligent agents. *International Journal of Human-Computer Studies*, *67*(1), 90-110.

81. Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250). ACM.

82. Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. Machine learning, 95(3), 423-469.

83. Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6), e98679.

84. Jansen, Y., Dragicevic, P., Isenberg, P., Alexander, J., Karnik, A., Kildal, J., ... & Hornbæk, K. (2015, April). Opportunities and Challenges for Data Physicalization. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15). ACM, New York, NY, USA, 3227-3236.

85. Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems* (pp. 325-333).

86. Ju, W., & Leifer, L. (2008). The design of implicit interactions: Making interactive systems less obnoxious. *Design Issues*, *24*(3), 72-84.

87. Ju, W., Lee, B. A., & Klemmer, S. R. (2008, November). Range: exploring implicit interaction through electronic whiteboard design. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (pp. 17-26). ACM.

88. Kammer, P. J., Bolcer, G. A., & Bergman, M. (1998). Requirements for Supporting Dynamic and Adaptive Workflow on the WWW. In *Proceedings of the Workshop on Adaptive Workflow Systems at CSCW'98*.

89. Kay, J., Kummerfeld, B., & Lauder, P. (2003, June). Managing private user models and shared personas. In *UM03 Workshop on User Modeling for Ubiquitous Computing* (pp. 1-11).

90. Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016, May). When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092-5103). ACM.

91. Kim, B. (2015a). Interactive and interpretable machine learning models for human machine collaboration (Doctoral dissertation, Massachusetts Institute of Technology).

92. Kim, B., Glassman, E., Johnson, B., & Shah, J. (2015b). iBCM: Interactive Bayesian Case Model Empowering Humans via Intuitive Interaction.

93. Kim, B., Malioutov D. M., Varshney, K. R. Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). *arXiv preprint*, arXiv:1607.02531.

94. Kim, B., Rudin, C., & Shah, J. A. (2014). The bayesian case model: A generative approach for case-based reasoning and prototype classification. In Advances in Neural Information Processing Systems (pp. 1952-1960).

95. Ko, A. J., & Myers, B. A. (2004, April). Designing the whyline: a debugging interface for asking questions about program behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 151-158). ACM.

96. Ko, A. J., & Myers, B. A. (2010). Extracting and answering why and why not questions about Java program output. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *20*(2), 4.

97. Koesten, L. M., Kacprzak, E., Tennison, J. F., & Simperl, E. (2017, May). The Trials and Tribulations of Working with Structured Data:-a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1277-1289). ACM.

98. Kokash, N., Birukou, A., & D'Andrea, V. (2007). Web service discovery based on past user experience. In *Business Information Systems* (pp. 95-107). Springer Berlin/Heidelberg.

99. Krause, J., Perer, A., & Ng, K. (2016, May). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686-5697). ACM.
100. Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012, May). Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1-10). ACM.
101. Kulesza, T., Wong, W. K., Stumpf, S., Perona, S., White, R., Burnett, M. M., ... & Ko, A. J. (2009, February). Fixing the program my computer learned: Barriers for end users, challenges for the machine. In *Proceedings of the 14th international conference on Intelligent user interfaces* (pp. 187-196). ACM.
102. Lacave, C., & Díez, F. J. (2002). A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, *17*(2), 107-127.
103. Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016, August). Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1675-1684). ACM.
104. Lambiotte, R., Delvenne, J. C., & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*.
105. Le Marc, M., Courtial, J. P., Senkovska, E. D., Petard, J. P., & Py, Y. (1991). The dynamics of research in the psychology of work from 1973 to 1987: From the study of companies to the study of professions. *Scientometrics*, 21(1), 69-86.
106. Lee, M. K., & Baykal, S. (2017, February). Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17). ACM, New York, NY, USA, 1035-1048.
107. Lee, M. K., Kim, J. T., & Lizarondo, L. (2017, May). A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3365-3376). ACM.
108. Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015, April). Working with machines: The impact of algorithmic and data-driven management on human workers. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1603-1612). ACM.
109. Lee, T. Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., & Findlater, L. (2017). The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105, 28-42. ISO 690.
110. Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2012). Building interpretable classifiers with rules using Bayesian analysis. *Department of Statistics Technical Report tr609, University of Washington*.
111. Lim, B. Y., Ayalon, O., & Toch, E. 2017. Reducing Communication Uncertainty with Social Intelligibility: Challenges and Opportunities. In *CHI 2017 Workshop on Designing for Uncertainty in HCI*.
112. Lim, B. Y., & Dey, A. K. (2009b, September). Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing* (pp. 195-204). ACM.
113. Lim, B. Y., & Dey, A. K. (2010, September). Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 13-22). ACM.
114. Lim, B. Y., & Dey, A. K. (2011, August). Design of an intelligible mobile context-aware application. In Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (pp. 157-166). ACM.
115. Lim, B. Y., & Dey, A. K. (2011, September). Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 415-424). ACM.
116. Lim, B. Y., & Dey, A. K. (2013, July). Evaluating Intelligibility Usage and Usefulness in a Context-Aware Application. In *International Conference on Human-Computer Interaction* (pp. 92-101). Springer, Berlin, Heidelberg.
117. Lim, B. Y., Dey, A. K., & Avrahami, D. (2009, April). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2119-2128). ACM.
118. Lim, B. Y., Smith, A., & Stumpf, S. (2018). ExSS: Explainable smart systems. In *Proceedings of the 23rd ACM International Conference Adjunct Papers on Intelligent User Interfaces.*
119. Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint*, arXiv:1606.03490.

120. Liu, Y., Goncalves, J., Ferreira, D., Hosio, S., & Kostakos, V. (2014, September). Identity crisis of ubicomp?: mapping 15 years of the field's development and paradigm change. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp '14). ACM, New York, NY, USA, 75-86.

121. Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014, April). CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14). ACM, New York, NY, USA, 3553-3562.

122. Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition*, *110*(2), 248-253.

123. Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303-332.

124. Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167-204.

125. Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *Research and Advanced Technology for Digital Libraries*, 473-474.

126. Madigan, D., Mosurski, K., & Almond, R. G. (1997). Graphical explanation in belief networks. *Journal of Computational and Graphical Statistics*, *6*(2), 160-181.

127. Malloch, J., Griggio, C. F., McGrenere, J., & Mackay, W. E. (2017, May). Fieldward and Pathward: Dynamic Guides for Defining Your Own Gestures. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 4266-4277). ACM.

128. Marquardt, N., Ballendat, T., Boring, S., Greenberg, S., & Hinckley, K. (2012, November). Gradual engagement: facilitating information exchange between digital devices as a function of proximity. In Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces (pp. 31-40). ACM

129. Marshall, J., Linehan, C., Spence, J. C., & Egglestone S. R. (2017). A Little Respect: Four Case Studies of HCI's Disregard for Other Disciplines. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17). ACM, New York, NY, USA, 848-857.

130. Mary Beth Kery, Amber Horvath, and Brad Myers. 2017. Variolite: Supporting Exploratory Programming by Data Scientists. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17). ACM, New York, NY, USA, 1265-1276.

131. Mennicken, S., Vermeulen, J., & Huang, E. M. (2014, September). From today's augmented houses to tomorrow's smart homes: new directions for home automation research. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 105-115). ACM.

132. Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv preprint*, arXiv: 1706.07269

133. Mozina, M., Demsar, J., Kattan, M., & Zupan, B. (2004, September). Nomograms for visualization of naive Bayesian classifier. In *PKDD* (Vol. 4, pp. 337-348).

134. Muñoz-Avila, H., & Aha, D. (2004). On the role of explanation for hierarchical case-based planning in real-time strategy games. In *Proceedings of ECCBR-04 Workshop on Explanations in CBR* (pp. 1-10).

135. Myers, B. A., Weitzman, D. A., Ko, A. J., & Chau, D. H. (2006, April). Answering why and why not questions in user interfaces. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 397-406). ACM.

136. Nikita, M. (2018). *Tuning of the Latent Dirichlet Allocation Models Parameters [R package ldatuning version 0.2.0]*. *Cran.r-project.org*. Retrieved 7 January 2018, from https://cran.r-project.org/web/packages/ldatuning/

137. Norman, D. (2009). *The design of future things*. Basic books.

138. Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic Books (AZ).

139. Núñez, H., Angulo, C., & Català, A. (2002, April). Rule extraction from support vector machines. In *Esann* (pp. 107-112).

140. Palmiter, S., & Elkerton, J. (1993). Animated demonstrations for learning procedural computer-based tasks. *Human-Computer Interaction*, *8*(3), 193-216.

141. Patel, K., Bancroft, N., Drucker, S. M., Fogarty, J., Ko, A. J., & Landay, J. (2010, October). Gestalt: integrated support for implementation and analysis in machine learning. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (pp. 37-46). ACM.

142. Patel, K., Drucker, S. M., Fogarty, J., Kapoor, A., & Tan, D. S. (2011, July). Using multiple models to understand data. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, No. 1, p. 1723).

143. Pierce, J., & Paulos, E. (2012). Beyond energy monitors: interaction, energy, and emerging energy systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 665-674.

144. Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D. S., ... & Anvik, J. (2006, July). Visual explanation of evidence with additive classifiers. In *Proceedings Of The National Conference On Artificial Intelligence* (Vol. 21, No. 2, p. 1822). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999

145. Raskar, R., Van Baar, J., Beardsley, P., Willwacher, T., Rao, S., & Forlines, C. (2006, July). iLamps: geometrically aware and self-configuring projectors. In *ACM SIGGRAPH 2006 Courses*(p. 7). ACM.

146. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). 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* (pp. 1135-1144). ACM.

147. Richards, N. M., & King, J. H. (2014). Big data and the future for privacy.

148. Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, *20*(5), 589-600.

149. Rosenthal, S., Selvaraj, S. P., & Veloso, M. M. (2016, July). Verbalization: Narration of Autonomous Robot Experience. In *IJCAI* (pp. 862-868).

150. Roth-Berghofer, T., Leake, D. B., & Cassens, J. (2012, July). Explanation-aware Computing ExaCt 2012. In *Proceedings of the Seventh International ExaCt workshop*.

151. Rukzio, E., Hamard, J., Noda, C., & De Luca, A. (2006, September). Visualization of uncertainty in context aware mobile applications. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services* (pp. 247-250). ACM.

152. Schilit, B., Adams, N., & Want, R. (1994, December). Context-aware computing applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on* (pp. 85-90). IEEE.

153. Schmidt, A., & Herrmann, T. (2017). 2017. Intervention user interfaces: a new interaction paradigm for automated systems. *interactions* 24, 5 (August 2017), 40-45.

154. *Scopus | The largest database of peer-reviewed literature | Elsevier*. (2018). *Elsevier.com*. Retrieved 7 January 2018, from https://www.elsevier.com/solutions/scopus

155. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. See https://arxiv. org/abs/1610.02391 v3.

156. Shneiderman, B. (2016). Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of Sciences*, *113*(48), 13538-13540.

157. Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents. *interactions*, *4*(6), 42-61.

158. Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2016). Grand challenges for HCI researchers. *interactions*, *23*(5), 24-25.

159. Sørmo, F., Cassens, J., & Aamodt, A. (2005). Explanation in case-based reasoning–perspectives and goals. *Artificial Intelligence Review*, *24*(2), 109-143.

160. Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., & Penney, D. (2016). Interpretable Machine Learning Models for the Digital Clock Drawing Test. *arXiv preprint arXiv:1606.07163*.

161. Srba, I., & Bieliková, M. (2010, August). Tracing strength of relationships in social networks. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 3, pp. 13-16). IEEE.

162. Stumpf, S., Rajaram, V., Li, L., Wong, W. K., Burnett, M., Dietterich, T., ... & Herlocker, J. (2009). Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, *67*(8), 639-662.

163. Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge university press.

164. Stolze, M., & Ströbel, M. (2003). Dealing with learning in ecommerce product navigation and decision support: the teaching salesman problem. In *Proceedings of the Second Interdisciplinary World Congress on Mass Customization and Personalization*.

165. Swartout, W. R. (1983). XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence*, *21*(3), 285-325.

166. Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.*, *11*, xxvii.

167. Thebault-Spieker, J., Terveen, L., & Hecht, B. (2017). Toward a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *24*(3), 21.

168. Trout, J. D. (2007). The psychology of scientific explanation. *Philosophy Compass*, *2*(3), 564-591.

169. Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007, April). How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 31-40). ACM.

170. Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: can it facilitate? *International Journal of Human-Computer Studies*, *57*(4), 247-262.

171. USACM. 2017. Statement on Algorithmic Transparency and Accountability. Accessed on 2017-09-09. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

172. van Melle, W., Shortliffe, E. H., & Buchanan, B. G. (1984). EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems. *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*, 302-313.

173. Vermeulen, J. (2010, September). Improving intelligibility and control in ubicomp. In *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing* (pp. 485-488). ACM.

174. Vermeulen, J. (2014). Designing for intelligibility and control in ubiquitous computing environments. (Doctoral dissertation). Hasselt University.

175. Vermeulen, J., Luyten, K., Coninx, K., & Marquardt, N. (2014, June). The design of slow-motion feedback. In *Proceedings of the 2014 conference on Designing interactive systems* (pp. 267-270). ACM.

176. Vermeulen, J., Luyten, K., van den Hoven, E., & Coninx, K. (2013, April). Crossing the bridge over Norman's Gulf of Execution: revealing feedforward's true identity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1931-1940). ACM.

177. Vermeulen, J., Slenders, J., Luyten, K., & Coninx, K. (2009, November). I bet you look good on the wall: Making the invisible computer visible. In *European Conference on Ambient Intelligence* (pp. 196-205). Springer, Berlin, Heidelberg.

178. Vermeulen, J., Vanderhulst, G., Luyten, K., & Coninx, K. (2009). Answering why and why not questions in ubiquitous computing. In *Ubicomp 2009 Posters*.

179. Vermeulen, J., Vanderhulst, G., Luyten, K., & Coninx, K. (2010, July). PervasiveCrystal: Asking and answering why and why not questions about pervasive computing applications. In *Intelligent Environments (IE), 2010 Sixth International Conference on* (pp. 271-276). IEEE.

180. Wensveen, S. A., Djajadiningrat, J. P., & Overbeeke, C. J. (2004, August). Interaction frogger: a design framework to couple action and function through feedback and feedforward. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* (pp. 177-184). ACM.

181. Wilson, A. G., Kim, B., & Herlands, W. Proceedings of NIPS 2016 Workshop on Interpretable Machine Learning for Complex Systems. *arXiv preprint*, arXiv: arXiv:1611.09139.

182. Wolff, S., & Grüter, B. (2008). Context, emergent game play and the mobile gamer as producer. In *GI Jahrestagung (1)*(pp. 495-500).

183. Yang, R., & Newman, M. W. (2013, September). 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* (pp. 93-102). ACM.

184. Youngblood, G. M., Cook, D. J., & Holder, L. B. (2005). Managing adaptive versatile environments. *Pervasive and Mobile Computing*, *1*(4), 373-403.

185. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus

186. Ziegler, C. N., Lausen, G., & Schmidt-Thieme, L. (2004, November). Taxonomy-driven computation of product recommendations. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 406-415). ACM.

187. Zilke, J. R., Mencía, E. L., & Janssen, F. (2016, October). DeepRED–Rule Extraction from Deep Neural Networks. In *International Conference on Discovery Science* (pp. 457-473). Springer International Publishing.