KEYWORDS:
ARTIFICIAL INTELLIGENCE; EXPLAINABLE AI; XAI, PROCESS THEORY; AI ETHICS

DOI:
https://doi.org/10.5771/2747-5174-2021-2-10

Ella Hafermalz *is Associate Professor at the KIN Center for Digital Innovation at Vrije Universiteit Amsterdam. Her research looks at new and old ways of working and organizing with digital technologies. Ella employs problematization and abductive methods to investigate emerging phenomena in an historically and contextually grounded way. Her research has been published in leading Information Systems and Organizational journals.*

Marleen Huysman *is professor of Knowledge and Organization at the School of Business and Economics of Vrije Universiteit Amsterdam where she leads the KIN research group and the KIN Center for Digital Innovation. She teaches and publishes on topics related to the practices of developing and using digital technologies - in particular artificial intelligence - and new ways of working. Her research has been published in various leading journals in the field of Information Systems and Organization Science.*

Please Explain: Key Questions for Explainable AI Research from an Organizational Perspective

AUTHORS: Ella Hafermalz & Marleen Huysman

ABSTRACT:

There is growing interest in explanations as an ethical and technical solution to the problem of ‚opaque' AI systems. In this essay we point out that technical and ethical approaches to Explainable AI (XAI) have different assumptions and aims. Further, the organizational perspective is missing from this discourse. In response we formulate key questions for explainable AI research from an organizational perspective: 1) Who is the ‚user' in Explainable AI? 2) What is the ‚purpose' of an explanation in Explainable AI? and 3) Where does an explanation ‚reside' in Explainable AI? Our aim is to prompt collaboration across disciplines working on Explainable AI.

1. INTRODUCTION

When Google maps tells you to turn right in 200 metres, you don't wonder "why" it is giving you that instruction. The application has delivered you to your destination in the past, and there is no reason to assume it will not do so again. However, if an algorithmic system denies your bank loan, or rejects your job application, you are likely to want an explanation. Yet often, there is no clear answer available. Not even an engineer can ask a deep learning algorithm: why did you do that?

Today even algorithms "know more than they can tell'. There is increasing awareness that this algorithmic "opacity" (Burrell, 2016) is problematic. So-called "black box" algorithms that work via deep learning draw on large sets of data to create their own models of reality in a way that generates remarkably accurate predictions. These models, for example in the case of neural nets, become so complicated as they optimize themselves, that even the scientists "in charge" cannot say exactly why the model can for example identify a picture of a cat with such high accuracy rates.

Black box algorithms can conceal biases that emerge from training data, for example when Amazon was forced to abandon its talent selection algorithm because it learned from past hiring patterns to discriminate against female applicants (Dastin, 2018). There is related concern that these systems are given too much autonomy, especially as they cannot be questioned about their actions in the case of a mistake being made. An extreme example is autonomous weapons mistakenly firing on a civilian - can the validity of the action be assessed when it is not possible to interrogate the rationale that underpinned it (Russell et al., 2015, Schulzke, 2013)?

Such scenarios are prompting a response from various communities, including regulators, ethicists, and computer scientists. All recognise that developments in Artificial Intelligence (AI) are unlikely to slow down, but that there is a need to ensure that these developments remain human-centric (Michal et al., 2009, Ohsawa and Tsumoto, 2006, Rosenberg, 2016), in line with social values, and to this end, explainable (Santiago and Escrig, 2017, Doran et al., 2017). While this conversation is multi-disciplinary, the work and organizational perspective is often missing from public discourse on how to make AI more responsible. This is surprising, given that organizations are a prime application context for new AI technologies (Faraj et al., 2018, Orlikowski, 2016, Lee, 2018). We aim to bring an organizational perspective to the Explainable AI (XAI) research agenda.

In this discussion paper, we show that the notion of "explanation" is emerging at the core of multi-disciplinary responses to the problem of opaque "black box" deep learning algorithms (Burrell, 2016). We critically inspect this notion of explanation as it appears in a much-discussed EU Commission text - The Ethics Guidelines on "Trustworthy AI", and in publicly available documents outlining a major research project funded by DARPA on "Explainable AI". Through our initial analysis of these texts and surrounding discourses, we show firstly that there is an apparent disconnect between an ethically motivated understanding of Explainable AI and a technically motivated one. We build on these points of disconnect by drawing on a processual and relational organizational perspective (Cecez-Kecmanovic et al., 2014) to begin to show how the invocation of the notion of explanation raises as-yet unresolved questions relevant to work and organizing.

In particular, we draw on a processual and relational perspective to ask: Who is the user of an explanation in the context of AI at work, and how does this relation change the nature of what an explanation is? And, relatedly, what is or could be the purpose of an explanation in these contexts, and what transformations do these varied purposes enact upon the nature of explanation? As these unresolved questions are identifiable as such from a relational and processual perspective, we wish to point to the possibility that scholars of work and organizing have to contribute productively to the explanation-driven response to "inscrutable" AI (Introna, 2016).

2. THE PROBLEM: OPAQUE "BLACK BOX" ALGORITHMS

The terminology around AI is often confused and can be misleading. For the purposes of this paper, we wish to emphasise a distinction between earlier forms of rule-based algorithms, featuring traceable decision trees and linear regressions, versus the new class of sophisticated "learning algorithms" (Faraj et al., 2018, O'Neil, 2017). These newer models are often trained on extremely large and sometimes unstructured data sets. The models can optimize themselves based on feedback as to whether the prediction they made was correct or not. A Deep Neural Network (DNN) for example learns a vast number of parameters without supervision, making networks of connections that do not make much sense to those who have the capacity to "peek under the hood". In this sense, the newest deep learning algorithms or models more closely resemble how humans learn to identify patterns and objects inductively. For example, you can quickly and confidently distinguish one friend from another, but could likely not give a very detailed account of how that distinction was arrived at in a split second (except through some form of post-hoc rationalisation, e.g. I could tell it was Emily because of her red hair and the way she walks).

Deep learning techniques are today remarkably accurate and powerful in their predictions. They already underpin many applications that people use in their everyday routines, such as facial recognition technology in photography applications and recommender systems on music and video streaming platforms (Dobbe et al., 2018). They are also increasingly used in business and societal contexts that are perhaps less obvious - for example in financial fraud detection and advising judiciary system outcomes such as sentencing and parole. Such uses are generating significant ethical concern and debate.

A central worry is that the evolving complexity of deep learning algorithms, how these algorithms work, is "inscrutable"[1] (Burrell, 2016) and they cannot therefore be held accountable for their recommendations (Introna, 2016). These algorithms therefore cannot easily be audited[2], or "held to account" (Roberts, 2009). A challenge therefore arises on several fronts: from an ethical, legal, and regulatory perspective deep learning algorithms are becoming suspect. In a connected way, the issue of inscrutability is also a problem for technical advancement: it is very difficult to "debug" a model that is not comprehensible. But the suspicion and mistrust that arises around the implementation of such systems also causes barriers to their use and acceptance in societal and organizational contexts.

While up until recently this commentary and critique has been oriented towards mostly social and legal options and solutions, such as calls for increased awareness and regulation, there is now a growing response from a more technical perspective. We group this technical response to the problem of opacity in learning algorithms under the umbrella term "Explainable AI".

Explainable AI has caught our interest for several reasons. Firstly, it is a proposed "solution" to the opacity problem that scholars of work and organization have pointed to in the context of learning algorithms in the era of digital work (e.g. Faraj et al., 2018, Orlikowski, 2016, Introna, 2016). Secondly, we saw that the idea of explainability is central to a new set of ethical guidelines for AI that has been produced for the European Commission (High Level Expert Group on AI Ethics Guidelines for Trustworthy AI, 2019), indicating that the broader ethics community also identify explanation as an important component of ethical AI in society. Thirdly, computer scientists are busy creating technical applications that are intended to make extremely complex models "human interpretable" through explanations. We are intrigued to know how this concept is operationalized in a field that holds quite different ontological and epistemological assumptions (Dobbe et al., 2018) from those that are held amongst in particular interpretivist scholars of work and organizational studies.

Finally, we recognise that a) in this highly multi-disciplinary conversation, the context of work and organizations is still largely missing, in favour of societal-level concerns, and b) the relevant insights and perspectives from theories and studies of work and organization, for example in relation to knowledge, learning, and professional expertise, are also mostly absent from the public discourse and technical initiatives directed at "solving" the problem

1 Intellectual Property rights are a further barrier to inspecting proprietary algorithms; see Edwards, L. and M. Veale (2017). "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for", Duke L. & Tech. Rev., 16, 18.

2 Though Cathy O'Neil, author of Weapons of Math Destruction (2017) has also begun a consulting initiative that audits algorithms, for example for KPMG; see O'Neil Risk Consulting & Algorithmic Auditing (ORCAA). Available at: http://www.oneilrisk.com/.

of opacity in learning algorithms. This is striking, given that many AI applications underpinned by learning algorithms are intended for use in organizational and work settings (Faraj et al., 2018).

Given this set of conditions, our approach so far has been exploratory. We have read academic articles from other fields such as ethics, law, and computer science, and have engaged with the broader public discourse surrounding the topic of Explainable AI. Our two main public points of entry to the Explainable AI conversation have been via the EU Commission's High Level Expert Group Guidelines on "Trustworthy AI" (in which explainability features strongly) and DARPA's large scale government funded project on Explainable AI (XAI), about which there is public information available online.

In the following we summarise our findings from our analysis of these sources and conversations. We then problematize several aspects of Explainable AI initiatives. To do so, we draw on a relational (Cecez-Kecmanovic et al., 2014) and processual (Langley et al., 2013) perspective. This means that we are sensitive to how phenomena are constituted through relationships, as well as their place in and across time. For example, we note that in the discourse on Explainable AI, the relationship between an explanation and a particular "user" is largely absent. A process perspective further adds a consideration of temporality, which we identify as lacking in Explainable AI's treatment of the concept of explanation. It is for example not clear "where" or "when" in a process an explanation is expected to reside - is an explanation meant to refer to pre-determined rules, a blow-by-blow account of what "actually" happened, or a post-hoc (re)construction? In the last section of this paper we use this analysis to motivate a set of questions that we hope can stimulate further inquiry into the role of explanation in learning algorithms in the context of work and organizations. The following analysis and suggested directions for further research are thus motivated by the question: What key questions can a relational process perspective in work and organizing bring to research on Explainable AI?

3. EXPLAINABLE AI: CURRENT APPROACHES

Explainability is seen by many as the key to making AI more transparent and therefore more trustworthy (e.g. Santiago and Escrig, 2017, Gunning, 2017, Doran et al., 2017). It is at first glance possible to therefore assume that diverse fields (e.g. Computer Science, Ethics, and Law) agree on what explainability means in the context of AI, what can be expected of this Explainable AI, and how it might work in practice. However, we have identified several gaps and points of confusion, if not outright contradiction, in how the idea of explainability is being applied in the context of AI. To begin to illustrate these points of disconnect, we compare how explainability is discussed in a recent set of Ethical Guidelines for Trustworthy AI commissioned by the EU with how explainability is discussed in documents relating to DARPA's project on Explainable AI (XAI), which started in 2017. These points of comparison are our initial analyses of the discourse surrounding Explainable AI.

First, we turn to the "Ethics Guidelines for Trustworthy AI" as a starting point for examining how explainability is discussed in the context of AI. These Guidelines were created by the 52 members of the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG). This Expert Group consists of computer scientists, ethicists, policy makers, business leaders, and others. They explain in their report that they "believe that AI has the potential to significantly transform society", and also that „AI is not an end in itself, but rather a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation" (High Level Expert Group on AI Ethics Guidelines for Trustworthy AI, 2019, p. 4). To achieve such benefits for society and for the environment, they argue that AI systems "need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom" (Ibid, p. 4, emphasis added).

The Guidelines have a broad scope, and emphasise a societal-level perspective. The notion of explainability[3] is one of several concepts that are central to the Ethics Guidelines, which

are aimed at businesses and public organisations that are developing and using AI in a way that impacts customers and citizens. Explainability is introduced and defined in this context as follows:

> **Explainability**. Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency). (High Level Expert Group on AI Ethics Guidelines for Trustworthy AI, 2019, p. 18)

We wish to draw particular attention to two demands that are stated here: firstly, a demand for (qualitative) explanations that are understandable by everyone from layperson to researcher and secondly, a demand for (quantitative) explanations assessing the degree to which an algorithm has "influenced and shaped organizational decision-making". We contend that scholars familiar with theories of decision making and knowledge management would be well placed to point out the enormous complexity involved in responding to such demands with accurate and timely explanations.

As we have already mentioned, explainability is particularly difficult in the context of deep learning algorithms, which are designed for superior prediction accuracy, not legibility. Therefore we can conclude that the guidelines that relate to explainability in AI are very ambitious, in both the nature of what they assume can be explained (neural nets; level of influence of one intervention on organizational processes), when these explanations can be offered (on demand, implying that they are always available), and to whom such explanations need to be intelligible (layperson, regulator, researcher). We will return to these points, but first we compare this high-level ethical discussion of explainability in AI with an ambitious technical project ostensibly concerned with the same idea: DARPA's XAI.

"Explainable Artificial Intelligence", also referred to as "XAI", is a growing research field in Computer Science. There are several groups of researchers dedicated to developing AI that offers some form of explanation for its outputs, however the most established and well known (and possibly best funded) is a DARPA initiative. DARPA is the agency of the United States Department of Defense that is responsible for developing emerging technologies for use by the military; it stands for The "Defense Advanced Research Projects Agency". The overall purpose of the DARPA XAI initiative is stated as part of the attempt to create "third-wave AI", a combination of early expert systems, and newer statistical forms of machine learning (Gunning, 2017).

DARPA's XAI project involved a call for proposals by teams who now work on different approaches to achieving explainability in AI. What exactly is meant by an explanation is left quite vague in many of their initial publicly available materials. A key concern within the project is to address a perceived "tradeoff" between explainability and performance[4], particularly in machine learning:

> The XAI program would like to improve explainability while maintaining a high level of learning performance for a range of machine learning techniques. There is an inherent tension between machine learning performance (predictive accuracy) and explainability; often the highest performing methods (e.g., deep learning) are the least explainable, and the most explainable (e.g., decision

3 It is sometimes used interchangeably with the word 'explicability', meaning the capacity to be able to be explained

4 Although the trade-off between performance and interpretability has been labelled a 'myth' in Rudin, C. (2018). "Please stop explaining black box models for high stakes decisions", arXiv preprint arXiv:1811.10154.

> trees) are less accurate. The program plans to fund a variety of machine learning techniques to provide future developers with a range of design options covering the performance versus explainability trade space. Explainable models might be created by learning to associate explanatory semantic information with features of the model, by learning simpler models that are easier to explain, by learning richer models that contain more explanatory content, or by inferring approximate models solely for the purpose of explanation. (DARPA, 2016, p. 7)

There are roughly three different approaches to Explainable AI captured in the above. The first strategy is to simplify a machine learning model to the point where the engineer can grasp all of its parameters and can point to how an outcome was arrived at. This option severely diminishes performance but has high explainability; a trade-off that few computer scientists are likely to be satisfied with. A second option is to create a more human-interpretable overlay that sits on top of an existing machine learning model. This requires a great deal of extra effort, and the result may only be somewhat interpretable to the human, depending on the interpreter's level of technical expertise. The third option involves decoupling the explanation from the original machine learning model altogether, and providing explanations only as a kind of "interface" that caters to a human desire for a rationalisation, but this rationalisation may have little to no connection to what is "really" going on inside a black box model.

From a technical perspective, the second two options in particular represent an exciting design and engineering challenge. However, in our reading thus far we notice little connection between the ideals of explainability as communicated in the Ethics Guidelines versus such technical discussions surrounding XAI. To further illustrate a potential disconnect, we show DARPA's depiction of the design challenge which will be addressed in their Explainable AI program:

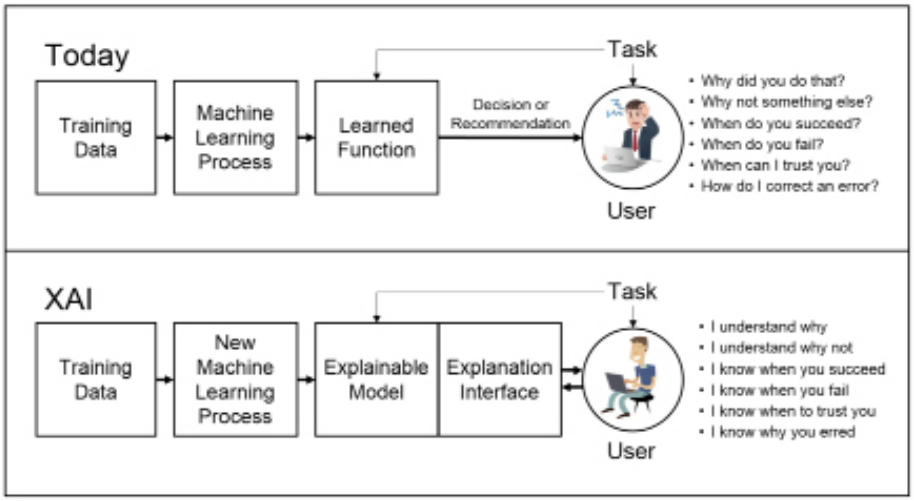

Figure 1: DARPA's XAI Concept (DARPA, 2016, p. 6)

In this image, "task" and "user" are prominent terms, which are perhaps to be expected in a Computer Science publication. We note however that the idea of "user" and "task" is difficult to map onto what was discussed in the Ethics Guidelines. For example, the Ethics Guidelines are concerned with AI systems that have "a significant impact on people's lives", for instance loan applications that are approved or denied autonomously by a learning algorithm. Who is the "user" in this example - the bank, or the customer? And what is the "task" - the loan application or its assessment? It is therefore ambiguous what use case the problem of XAI applies to, and who the system's explanations might be aimed at.

The nature of an explanation from the DARPA initiative's perspective is operationalised in the statements listed in the bottom right hand corner of Figure 1, for example "I know when you fail" and "I know when to trust you". The Ethics Guidelines however referred also to the importance of explanations relating to how an AI system has influenced and shaped

an „organisational decision-making process". That type of explanation is not, as far as we are aware, currently considered in the scope of technical approaches to Explainable AI. Technical projects instead concentrate on solving the problem of explainability for a generic individual. On the other end, ethics discussions tend to be concerned with society. A result of both these foci is that the level of social practices in organizations is missed.

In this short overview we have introduced several points of potential tension within the Explainable AI discussion that we have begun to problematize (Alvesson and Kärreman, 2007, Alvesson and Sandberg, 2013). Our intention is not merely to criticise these quite recent and ongoing initiatives. Indeed, we recognise that this summary is from the perspective of "outsiders" to the fields under study (ethics and computer science). However as informed outsiders we are also well placed to see that there may be conceptual and practical fault-lines emerging, where two communities that are apparently united by a similar goal (opening the black box of learning algorithms through Explainable AI) may actually be talking past one another. We further notice that the organizational perspective and context is largely missing and begin to consider what this perspective could bring to the table.

Specifically, we identify three points of potential discrepancy or confusion that we will give further analytical attention in this paper from a processual, organizational perspective. We summarise these points as questions: Who is the user of an explanation in Explainable AI, and what difference does this make for the nature of the explanation? For what purposes could an explanation from AI be useful? And Where and therefore when in time does an explanation reside in Explainable AI? We now consider in greater detail what these questions mean and what reflections they prompt.

4. BRINGING AN ORGANIZATIONAL PERSPECTIVE TO XAI RESEARCH

There are several ways that Explainable AI is relevant for organisational and Information Systems scholars, and vice versa. Explainable AI development is most obviously an "object" of study (digital innovation, science in action, emerging technologies). Organizational scholars could alternatively enter the fray of the ethical discussions that are surrounding AI's impact in society, where Explainable AI is emerging as a hopeful initiative for those who warn of the deleterious effects of inscrutable, black boxed algorithms. More ambitiously, scholars of work and organizing could and arguably should aim to bring our own expertise to these other actors (computer scientists, law scholars and regulators, etc) who are already shaping the discussion on explanation in AI, and may have blind spots that can be addressed from processual perspectives on knowledge and learning, professional expertise, and working and organizing with digital technologies (e.g. Waardenburg, Huysman, and Sergeeva, 2021). This latter ambition is what we are working towards here.

This section is thus motivated both by our initial problematization of discussions that we have so far introduced, and by new questions that have emerged as we have tried to take seriously the ambition and aims of those who are pursuing Explainable AI. The following three questions act as a starting point for contributing to key concepts in Explainable AI with expertise from the fields of working and organizing with a relational and processual perspective in mind. The aim is that by sketching out these areas for further discussion we also take a first step to bridging disciplinary boundaries, a necessary exercise for the purposes of ensuring that the future of AI in work and organizations becomes more rather than less human-centric.

4.1. WHO IS THE "USER" IN EXPLAINABLE AI?

A first question that arises from a relational organizational perspective concerns who the "user" is in Explainable AI. In discourse on Explainable AI, the "user" is hardly defined. Of course, the inadequacy of the very term "user" and the complex ways in which information systems and technologies themselves even "configure the user" (Woolgar, 1990, Suchman,

2007) has long been acknowledged in Human Computer Interaction and Information Systems research. This richer understanding of the user does not seem to be given much attention in the discourse around the Explainable AI initiative thus far. In the Ethics Guidelines for example, we identified reference to a layperson, regulator, or researcher as a potential "user" of an explanation that relates to AI. However in DARPA's Figure 1 diagram, we see a single non-specific "user" scratching his head as he tries to make sense of the output he has received from a system assisting him with a task. These examples of who a "user" is represent quite different ways of thinking about who an explanation is for in the context of Explainable AI.

Further, none of these users (layperson/regulator/researcher/generic user) sheds light on what happens when AI is introduced in an organizational setting. We believe that this question of who the user is, is particularly relevant in the context of AI at work, because learning algorithms do not reflect "normal" support technologies that assist a "user" with completing a "task" (Faraj et al., 2018). Instead, some even believe that the latest generation of AI applications will replace the people whose expertise they emulate (Susskind and Susskind, 2015). Who is the user, then?

We will illustrate our point here by invoking different possible types of "user" of AI in an organizational context. Considering these four potential "explanation users" allows us to generate questions about the relational nature of explanations. Experts do not explain a concept, for example 'gravity', the same way to different audiences (e.g. a child versus a physics student), and we suspect that Explainable AI will need to gain an appreciation for the relational dimension of explanation before the explanations that are offered can become truly beneficial to their specific "user".

Is the user the engineer? This is a person who is responsible for developing and adjusting the performance of the AI system. Perhaps they would be most interested in a technical explanation that allows them to adjust any obvious problems that are reducing the accuracy rate of predictions. These explanations need to be technical in nature, and are unlikely to be "interpretable" by a "layperson'."

Is the user a domain expert? For example a radiologist. A radiologist has expertise in identifying patterns in an image, that they can interpret as indicating a cancerous growth. If an AI system has been created by an engineer, trained on a large data set, and has been found to predict cancerous growths in x-rays with an accuracy rate that is comparable to the radiologist - is the radiologist a user-candidate? What kind of explanations would be useful to her? Here we identify a crucial dimension of the eventual use of Explainable AI that is not yet adequately addressed: is the "user" at risk of being replaced by the system? If so, what do they stand to gain from an explanation? Or, does the presence of an explanation change the usual "automation" story, so that the expert (e.g. the radiologist) can instead gain a new kind of peer, to "bounce ideas off"?

Is the user a decision subject? Take the example of AI used to assess a loan application, and it is denied. Here, an explanation could be useful to the "decision subject" as a user, especially if this explanation can help the customer to understand whether or what they can do to change the outcome in a future scenario, for example by paying down credit card debt (Edwards and Veale, 2017). In an organizational setting, a job applicant who has been denied a position based on a machine-learning based screening process could similarly be termed a decision subject. If we count these parties as "users" of Explainable AI, what kind of explanations are needed? And are you still a "user" if you don't even realise that you are being impacted by an AI application's predictions, as is often the case in large-scale automated assessment and filtering scenarios?

Is the user a low-skilled worker? In the case of the platform economy, "algorithmic management" is already occurring. When an algorithm determines your driving routes, performance ratings, and even compensation rates (e.g. via Uber's surge pricing algorithm), are you still a "user" of this system? What kind of explanations are useful to a worker who is managed by AI? What can they do about it, if they disagree?

The notion of there being a "user" of explanations is, we suggest, in itself a productive

notion. The relationality of explanation in the context of XAI requires, we argue, a better appreciation of the nuanced roles that emerge in relation to AI, particularly in the workplace. The above illustrations are intended as a starting point for building up a relational perspective on the concept of the user in the context of Explainable AI.

4.2. WHAT IS THE "PURPOSE" OF AN EXPLANATION IN EXPLAINABLE AI?

Once we think about the various "users" of an explanation, we quickly come to appreciate that there are different purposes for an explanation. We begin by outlining four possibilities for examining the purpose of explanations below. Looking at organisational contexts and cases, as well as examining existing theoretical frameworks in organizational theory, could help in furthering the understanding of this issue through knowledge exchange across fields. Is the purpose of an explanation accountability? Much of the Explainable AI conversation and research is driven by the well-justified ethical need to hold AI systems to account for the at times disproportionate, untested, and potentially biased impact that they are having on citizens, for example in the case of informing prison sentencing and parole decisions. It is clear that action needs to be taken in response to unjust actions, however it has not been widely shown that increased demands for accountability necessarily lead to more ethical practices or indeed greater insight into the inner workings of organisation (Roberts, 2009). Further, even if an explanation can identify how a decision is made, it may not satisfy what the user actually wants - for example for the opportunity to change the outcome, or to receive compensation (Edwards and Veale, 2017).

Is the purpose of an explanations technology acceptance? Perhaps implicit in Computer Science literature on Explainable AI is an economically and politically motivated desire for exciting and powerful new technologies to be accepted by users, organisations, and society. If AI is "trustworthy", it will be used. Explanation and its assumed role in trust then becomes a way of comforting users (or those impacted by AI systems). Explanation may make a system seem less frightening, less "other", and thus more amenable to adoption and appropriation. The notion of "ethics-washing" (Wagner, 2018) may be a useful way of contextualising this strategic use of explanations in the context of AI.

Is the purpose of an explanation learning? If AI systems are indeed set to become expert "colleagues" in the future, perhaps the purpose of explanation is that employees can learn from the reasoning of their new "digital peers" - "Why did you select this candidate and not that candidate?" The reasoning that is offered in response to such a question could for example prompt an HR professional to question their own established heuristics for "what makes a good candidate". What do social learning and knowledge experts have to say about the role of explanation in learning and in the formation of communities of practice (Brown and Duguid, 2017)? What makes a "good" explanation in an organizational context, and when is it needed? Could explanations be the key to learning from AI-as-peer?

Is the purpose of an explanation collaboration? Explanations may play an important role in enabling seamless collaboration between human and non-human agents. In User Interface design for example, interfaces are being designed for recommender systems, where each recommendation is "inspectable" by clicking through to increasingly more detailed information about how the recommendation was arrived at (Shapiro, 2018). This is referred to as "progressive disclosure" (Nakatani and Rohrlich, 1983). These explanations are nested and can be uncovered at a rate and level of detail that is perceived as being useful by the user. Perhaps this way of examining rationales is already a part of how we work with colleagues - we can ask follow-up questions if we want to know more about a decision or judgement, but do not always require a full report on each team mate's behaviour. Such a gradual-release form of explanation could play an important role in Explainable AI's capacity to become a part of practice as a collaborator that can be queried, but that does not overwhelm the situation with irrelevant information.

4.3. WHERE AND WHEN DOES AN EXPLANATION "RESIDE" IN EXPLAINABLE AI?

Thirdly, the notions of explanation and explainability have an interesting relationship to temporality. Intuitively, explanations are "post-hoc" and yet are anchored to a "real" occurrence that has passed. For example, if a data breach occurs, a company is expected to explain why this breach occurred, who has been impacted, and what is being done to rectify the situation. Already such an explanation has multi-temporal foci - what occurred in the past, who is impacted now, and what will be done in the future. A sensitivity to temporality thus appears to be relevant when considering how explanation might work in the context of AI. We suggest that it is therefore relevant to consider, from a process-sensitive perspective, where and when in time an explanation resides.

Does an explanation reside in the setup of a system? In early AI systems, explanation could be said to lie latent in their setup. Earlier forms of AI are rule based and so the parameters of an algorithm are set up by a human before a computation takes place. If an explanation is requested of such a rule-based system, the initial set up can be inspected in order to formulate the explanation. Such a technical arrangement can be compared to the organizational structure of a bureaucracy, where formalised processes are recorded prior to organizational decisions being made. If a problem arises in relation to a decision, policies and rules are consulted to find an explanation for what has occurred and to decide on recourse. Learning algorithms and in particular deep learning techniques however are very different. They are more similar to a start-up organisation with agile ways of working that is rapidly growing: it is difficult to identify how and by whom a critical decision is or will be made, or what its consequences are or will be at an organisational level. In the case of learning algorithms, explanations cannot be seen as lying latent only in the "setup" of the system.

Does an explanation exist alongside a system? The abstract noun "explainability" implies that an explanation exists in a constant state, as a property of an entity. This conceptualisation perhaps best fits an analogy of firms and professions that are entrenched in "audit culture' (Satava et al., 2006), where every action is tied to a reporting of that action. For example, doctors and lawyers are tasked with constantly both performing their professional duties and also keeping records of their activities and decisions. The field of critical accounting in particular has much to say about the perverse effects of such a requirement: for example in the way that record-keeping can end up taking more time and resources than acting (Schwartz, 2014, Harding et al., 2010). This harks back to the performance/explainability trade-off mentioned earlier in the context of XAI, as the cost of creating a system around deep learning models, that records how a particular outcome was arrived at, is enormous (Edwards and Veale, 2017, Doshi-Velez et al., 2017).

Does an explanation exist only when called upon, or is it transformative of the whole process? Even when explanations are thought of as residing at the "end" of a process (for example after a deep learning algorithm has computed an output and when an explanation is requested), the entire process of machine learning can be transformed in anticipation of a post-hoc request to account for its actions. An important insight offered by Tsoukas (1997) and Roberts (2009) is that requests for transparency in the form of explanations (to give an account) can have performative effects. The desire for transparency does not just "open" a black boxed process, it transforms it. As Roberts (2009) details, the request to give an "account of oneself" requires a person to come up with a response that they believe will be accepted by the requestor. This means that calls for accountability require a performance of rationality that may not draw in a meaningful way on what "actually" happened but that nevertheless has significant consequences. We already see a practical instantiation of this performativity principle in initiatives that aim to create explanations for deep learning algorithms that are de-coupled from the actual inner workings of the model. To extrapolate from this, will a system become better at concealing its "true" actions, as it learns to better produce rationalisations that are approved of by a human audience? Performative effects are not straightforwardly good or bad, but they do raise questions for how explainable AI will be conceived of and put into practice.

CONCLUSION

Scholars of work and organizing are already discussing how AI will change work. There is particular concern regarding the problem of inscrutable, black box AI (Introna, 2016, Faraj et al., 2018, Orlikowski, 2016). These discussions are however also running, in parallel form, in neighbouring fields of ethics and computer science. Are we content to live in parallel universes, or is there a way to share insights and challenges when it comes to "solving" the problem of black box AI? While mutual contribution is our ultimate aim, we recognise that the first step is to know what is happening in other fields that are on the front lines of finding solutions to the opaque systems that are increasing ruling our lives and work (O'Neil, 2017).

We have shown that there is an active community of scholars and practitioners working at creating more transparent, „Explainable AI". Explainability is however an ideal that possesses much the same lure as the notion of transparency, and is perhaps as slippery and potentially paradoxical (Edwards and Veale, 2017). More information does not always lead to more understanding (Tsoukas, 1997). Yet currently the Explainable AI initiative seems generally to discount the relationality of explanations - for example, who is the "user" in the context of AI applications? And what is the purpose of an explanation, particularly in various organizational scenarios? Further, the temporality and correspondent performative effects of explanations are not yet being fully explored, yet we can begin to anticipate how systems that ostensibly add explanation "to the outer layer" as a form of accountability will become much more fundamentally transformed (Roberts, 2009). These are issues that an organizational perspective is sensitive to and we have used this stance to put forward key questions as prompts to future Explainable AI research.

Finally, the ethics discussion on Explainable AI often takes society and citizens as its focus, yet there is much to be said about how AI will impact organizational and work life. In drawing attention to an organizational perspective on Explainable AI we urge discussion about its merits, weaknesses, and our potential contribution to it. We hope to open an avenue for constructive research into how a processual and relational perspective on work and organizing can contribute to guiding a more positive future for AI at work.

ACKNOWLEDGEMENTS:

We are grateful for the feedback we received on this work at the 11th International Process Symposium in Crete, as well as to colleagues at the KIN Centre for Digital Innovation for their helpful comments on earlier versions of this paper.

REFERENCES:

Alvesson, M. and D. Kärreman (2007). "Constructing Mystery: Empirical Matters in Theory Development", *The Academy of Management Review*, 32, 1265-1281.

Alvesson, M. and J. Sandberg (2013). *Constructing research questions: Doing interesting research*, SAGE, London.

Brown, J. S. and P. Duguid (2017). *The social life of information: Updated, with a new preface,* Harvard Business Review Press.

Burrell, J. (2016). "How the machine "thinks": Understanding opacity in machine learning algorithms", Big Data & Society, *3, 2053951715622512.*

Cecez-Kecmanovic, D., R. D. Galliers, O. Henfridsson, S. Newell and R. Vidgen (2014). "The Sociomateriality of Information Systems: Current Status, Future Directions", *MIS Quarterly*, 38, 809-830.

DARPA, Agency, D.A.R.P. (2016) *Broad Agency Announcement: Explainable Artificial Intelligence* (XAI). Arlington, VA.

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*, Available: Reuters. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (Accessed 20 May 2019).

Dobbe, R., S. Dean, T. Gilbert and N. Kohli (2018). "A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics", *arXiv preprint* arXiv:1807.00553.

Doran, D., S. Schulz and T. R. Besold (2017). "What does explainable AI really mean? A new conceptualization of perspectives", *arXiv preprint* arXiv:1710.00794.

Doshi-Velez, F., M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger and A. Wood (2017). "Accountability of AI under the law: The role of explanation", *arXiv preprint* arXiv:1711.01134.

Edwards, L. and M. Veale (2017). "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for", Duke *L. & Tech. Rev.*, 16, 18.

Faraj, S., S. Pachidi and K. Sayegh (2018). "Working and organizing in the age of the learning algorithm", *Information and Organization*, 28, 62-70.

Gunning, D. (2017). *Explainable Artificial Intelligence (XAI): Defense Advanced Research Projects Agency.* Available at: https://www.darpa.mil/program/explainable-artificial-intelligence (Accessed: May 18 2019 2019).

Harding, N., J. Ford and B. Gough (2010). "Accounting for ourselves: are academics exploited workers?", *Critical Perspectives on Accounting*, 21, 159-168.

High Level Expert Group on AI Ethics Guidelines for Trustworthy AI (2019): The European Commission.

Introna, L. D. (2016). "Algorithms, governance, and governmentality: On governing academic writing", *Science, Technology, & Human Values*, 41, 17-49.

Langley, A., C. Smallman, H. Tsoukas and A. H. Van de Ven (2013). "Process studies of change in organization and management: unveiling temporality, activity, and flow", *Academy of Management Journal, 56, 1-13.*

Lauren Waardenburg , Marleen Huysman , Anastasia V. Sergeeva (2021) In the Land of the Blind, *the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. Organization Science* 0(0). https://doi-org.vu-nl.idm.oclc.org/10.1287/orsc.2021.1544

Lee, M. K. *(2018).* "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management", *Big Data & Society*, 5, 2053951718756684.

Michal, P., D. Pawel, S. Wenhan, R. Rafal and A. Kenji (2009). "Towards context aware emotional intelligence in machines: computing contextual appropriateness of affective states", *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence* (IJCAI-09), pp. 1469-1474.

Nakatani, L. H. and J. A. Rohrlich (1983). "Soft machines: A philosophy of user-computer interface design", In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 19-23.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*, Broadway Books.

O'Neil Risk Consulting & Algorithmic Auditing (ORCAA). Available at: http://www.oneilrisk.com/.

Ohsawa, Y. and S. Tsumoto (2006). *Chance discoveries in real world decision making: data-based interaction of human intelligence and artificial intelligence,* Springer.

Orlikowski, W. J. (2016). "Digital work: a research agenda".

Roberts, J. (2009). "No one is perfect: The limits of transparency and an ethic for "intelligent" accountability", *Accounting, Organizations and Society*, 34, 957-970.

Rosenberg, L. (2016). "Artificial Swarm Intelligence, a Human-in-the-loop approach to AI", In AAAI, pp. 4381-4382.

Rudin, C. (2018). "Please stop explaining black box models for high stakes decisions", *arXiv preprint* arXiv:1811.10154.

Russell, S., S. Hauert, R. Altman and M. Veloso (2015). "Ethics of artificial intelligence", *Nature*, 521, 415-416.

Santiago, D. and T. Escrig (2017) Why explainable AI must be central to responsible AI: Accenture. Available at: https://www.accenture.com/us-en/blogs/blogs-why-explainable-ai-must-central-responsible-ai (Accessed: 1/6/2019 2019).

Satava, D., C. Caldwell and L. Richards (2006). "Ethics and the auditing culture: Rethinking the foundation of accounting and auditing", *Journal of Business Ethics*, 64, 271-284.

Schulzke, M. (2013). "Autonomous weapons and distributed responsibility", *Philosophy & Technology,* 26, 203-219.

Schwartz, D. G. (2014). "The disciplines of information: Lessons from the history of the discipline of medicine", *Information Systems Research,* 25, 205-221.

Shapiro, V. (2018) ‚Explaining System Intelligence', SAP User Experience Community. Available at: https://experience.sap.com/skillup/explaining-system-intelligence/.

Suchman, L. A. (2007). *Human-machine reconfigurati ons : plans and situated actions,* Cambridge University Press, Cambridge ; New York.

Susskind, R. E. and D. Susskind (2015). *The future of the professions: How technology will transform the work of human experts*, Oxford University Press, USA.

Tsoukas, H. (1997). "The tyranny of light: The temptations and the paradoxes of the information society", *Futures*, 29, 827-843.

Wagner, B. (2018). "Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?", Being Profiling. Cogitas Ergo Sum.

Woolgar, S. (1990). "Configuring the user: the case of usability trials", T*he Sociological Review,* 38, 58-99.