AI & SOCIETY
https://doi.org/10.1007/s00146-023-01633-0

OPEN FORUM

Creating meaningful work in the age of AI: explainable AI, explainability, and why it matters to organizational designers

Kristin Wulff[1,2] · **Hanne Finnestrand**[1]

Received: 25 June 2022 / Accepted: 17 January 2023

Abstract
In this paper, we contribute to research on enterprise artificial intelligence (AI), specifically to organizations improving the customer experiences and their internal processes through using the type of AI called machine learning (ML). Many organizations are struggling to get enough value from their AI efforts, and part of this is related to the area of explainability. The need for explainability is especially high in what is called black-box ML models, where decisions are made without anyone understanding how an AI reached a particular decision. This opaqueness creates a user need for explanations. Therefore, researchers and designers create different versions of so-called eXplainable AI (XAI). However, the demands for XAI can reduce the accuracy of the predictions the AI makes, which can reduce the perceived usefulness of the AI solution, which, in turn, reduces the interest in designing the organizational task structure to benefit from the AI solution. Therefore, it is important to ensure that the need for XAI is as low as possible. In this paper, we demonstrate how to achieve this by optimizing the task structure according to sociotechnical systems design principles. Our theoretical contribution is to the underexplored field of the intersection of AI design and organizational design. We find that explainability goals can be divided into two groups, pattern goals and experience goals, and that this division is helpful when defining the design process and the task structure that the AI solution will be used in. Our practical contribution is for AI designers who include organizational designers in their teams, and for organizational designers who answer that challenge.

Keywords Artificial intelligence · Machine learning · Explainability · Sociotechnical systems (STS) · Organizational design

1 Introduction: the need for explainability

In today's development of AI solutions, effort and money are being wasted on creating solutions that never achieve their intended goals. One of the problems leading to this is AI solution's lack of explainability (Barredo Arrieta et al. 2020). The issue of explainability can be illustrated by two cases: the Pinot case and the loan case. In the first case, each summer, the software consultancy Pinot arranges for students to run a summer project where 300–400 students apply, with rising numbers every year. To deal with this, the artificial intelligence (AI) group at Pinot decided to create an AI solution, called StudRec, that classifies the students into two groups: *employ* or *not employ*. The AI team considered two machine learning (ML) models to be good choices: a simple decision tree or a neural network. The simple decision tree is more transparent, making it easier to understand the AI's classification. The neural network provides a result with a higher probability of correct classification but is more opaque and harder to understand. The student summer project is an important recruitment channel for Pinot, so the CEO was concerned about the selection outcome. The team explained that it could either use a less accurate and more transparent ML model, the simple decision tree, or the neural network model and add technology that explains what is going on: an eXplainable AI (XAI). In this first case, the CEO was involved in the AI team's deliberations. In a second case, described by Strich et al. (2021), another organization and its CEO were seemingly not involved in how their loan consultants' new AI-assisted way of working was being designed. In this case, the company reduced its loan consultants' responsibility from evaluating applicants and deciding

✉ Kristin Wulff
kristin.wulff@kantega.no

[1] Department of Industrial Economics and Technology Management, Faculty of Economics and Management, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[2] Kantega AS, Trondheim, Norway

Published online: 26 January 2023

who would get a loan, to them entering client information into an AI solution and delivering the AI's decision to the client. By doing so, it not only harmed the work identity of the loan consultants but also designed a task structure which increased the need for explanations on how the decision had been arrived at.

What we have described above is the problem of explainability. The requirements for explainability can come from the context (Lawless et al. 2019), the need for accountability (Kim and Doshi-Velez 2021), and the needs of the domain which is going to use the AI solution (Leslie 2019). Explainability is often seen from a merely instrumental perspective, but it is also intrinsically valuable in maintaining people's dignity, control over processes of decision-making, and self-actualization (Colaner 2022). Explainability is not a new challenge in AI development (Sørmo et al. 2005). What is new is the increased opacity of how the AI reached its decisions introduced by the use of ML, and especially deep learning (Barredo Arrieta et al. 2020). In most cases, not even the designer of the AI solution can explain why it reached the decision it did. However, to ascertain that the AI solution produces results that are fair, trustworthy, and consistent, there is a need for the humans involved to understand what is happening, hence the efforts at explainability.

The focus of explainability research has been on *technology*, concentrating on how the XAI is able to explain anything (Gunning 2016), on the *people* involved, examining the user roles (also called audiences, Barredo Arrieta et al. 2020; Rossi 2019) and prescribing a user-centric explainability (Liao et al. 2020), on how the *explanations* are formulated (Miller 2019), and on the *interaction* between technology and people (Bussone et al. 2015). There is also research into the *design process* for the AI solution, for instance by providing design guidelines (Eiband et al. 2018), and creating design methods like scenario-based explainability design (Wolf 2019). These are important perspectives, but there is also a need to see this research in the context of organizations' task structure design, because how the task structure is designed sets guidelines for the use of AI. Unfortunately, many organizations have a task structure design that is counterproductive to the demands of the environment (de Sitter et al. 1997). This can influence the introduction of AI because the explainability demands are based on non-optimal organizational designs that create a higher need for explanations. This can create exaggerated expectations of the XAI that the data scientist will not be able to fulfill, which can lead to the AI solution being abandoned before reaching the production stage or to choosing non-optimal ML models. Those organizations that are likely to succeed in their use of AI examine their task structures and redesign them where necessary (Barro and Davenport 2019). The design of the task structure often begins with a top-down view of the tasks that the organization will need to perform to serve its customers (users), and an exploration on how to best divide up the tasks (de Sitter et al. 1997; Worren 2018). To further explain the different design options, it is expedient to make use of the design principles from socio-technical systems design (STSD). STSD aims to enhance the performance of work systems by acknowledging that leveraging the knowledge and capabilities of workers with the assistance of technological systems will achieve better operational performance when dealing with uncertainty, variation, and adaptation (Pasmore et al. 2019). STSD has gained some renewed interest when studying digitalization due to the introduction of new work systems in organizations (Babüroğlu and Selsky 2021; Guest et al. 2022). However, STSD can be useful when designing IT solutions for other types of organizations as well (Govers and Südmeier 2016), where it is crucial to handle the complexity of the environment and prevent the IT solution from making it more rigid (Govers and van Amelsvoort 2019). The two main patterns in organizational design are either to divide the task structure into small units of work, as, for instance, in factory lines where each person performs just one small task, or to give responsibility for a larger part of the task structure to a set of people (group/team), as, for instance, within an emergency response team. Whether to go with the first or second pattern depends on how much variety there is in the external or internal work environment (Emery and Trist 1965). The more variety, the more advisable it is to give the responsibility for a larger part of the task structure to an operational work team (de Sitter et al. 1997). STSD's other concern is with the design of the control structure, which is ideally bottom up, meaning that any problems that can be solved by the employee/team should be solved by them (de Sitter et al. 1997).

Because AI solutions can cause an increased need for explainability, there is a need to disentangle the explainability needs that are caused by a non-optimal organizational design from the true explainability needs. This means that the design team that is given the task of increasing the explainability of an AI solution will benefit from understanding the options of how a task structure can be designed. For the organization, it will be beneficial to optimize the task structure for their environment regardless of the use of AI. Therefore, it is recommended to optimize the task structure before, or at least as a part of, the introduction of the AI solution. This leads us to the question:

1.1 How can organizational task structure be designed to reduce or change the need for XAI?

In Sect. 2 of this paper, we present the explainability challenges presented by the use of ML, before describing STSD and its design principles and parameters in Sect. 3. The actors participating in the AI design and development are

presented in Sect. 4, where we suggest a more interdisciplinary team than what may be the standard today. In Sect. 5, we discuss and present the benefits and dangers of different organizational designs vis-à-vis the explainability necessary in an organization, before concluding in Sect. 6 and suggesting further research.

2 Machine learning and explainability

Machine learning (ML) is "a subfield of AI that studies the ability to improve performance based on experience" (Russell and Norvig 2021, p. 1). ML models come in many variants with differing levels of opaqueness. This affects the explainability of the ML. Some ML models are self-explanatory. These are called white-box learning models (Hall 2018), as opposed to ML models described as black boxes (Buhrmester et al. 2021; Castelvecchi 2016; Faraj et al. 2018). White-box ML are, for instance, tools like small regression models and decision trees. In regression models, one can have a list of the variables that are relevant for predicting what one wants to predict, and a weight for how much they affect the resulting decision. In a small decision tree, one can see the decisions on the leaf node level, trace the data through the tree, and see which nodes have been relevant for the outcome (Russell and Norvig 2021).

In white-box ML models, the designers can see what is going on. In black-box ML, however, there can be deep neural networks that work by shifting the weight on nodes in several layers as the neural network learns (Guidotti et al. 2018). Here, it is not possible to fully understand what is going on. One could print out the documentation, but it is so extensive that it is not humanly possible to comprehend. In the case of some of these AI solutions, it is easy to evaluate whether the result from the AI solution is correct or not, for instance in image recognition training, where the task is to recognize certain objects in an image. Other types of results, for instance interpreting lab values in medical research, are harder for humans to confirm. In such cases, one must find ways of communicating the results so that the humans involved can evaluate them, by, for example, creating an output that clinicians can understand and use in their work (Shao et al. 2021). Another way is to add post-hoc analysis (Russell and Norvig 2021). Such analysis can take two routes: either manipulating the data going into the AI solution to see how that changes the results, or analyzing a component inside the complex ML model (Liang et al. 2021). The design team can also run ML models with different transparencies in parallel and check that the results correlate to satisfaction, for instance by running a deep neural network and a simple and explainable ML model in parallel (Shao et al. 2021).

Explainability of AI solutions and the design of the appropriate XAI have been studied increasingly over the last years (Minh et al. 2022). The term XAI was coined by Van Lent et al. (2004) to describe the ability of the AI solution to give an explanation either during or after execution of the task. A definition of XAI is that the technology explains to the people involved how to understand each decision, recommendation, or action and the process followed to reach it (Biran and Cotton 2017). The difficulty of finding ways to making things understandable is evident in the fact that explainability is one of the main barriers to implementing AI solutions today (Barredo Arrieta et al. 2020). The importance of effectively handling the issue of explainability is evident in the goals that explainability aims to fulfill. In their literature review on explainability, Barredo Arrieta et al. (2020) found nine goals of explainability: *fairness*, *privacy awareness, causality, transferability, accessibility, confidence, informativeness, interactivity,* and *trustworthiness.*

Fairness is perhaps the most important explainability goal (Benbya et al. 2020) because the AI solution might develop in a way that makes the technology perform unjust actions and reinforce biases (d'Alessandro et al. 2017; Escalante et al. 2018). Therefore, research into XAI is important from an ethical point of view (Heinrichs 2022). For example, in the Pinot case, the certainty that the StudRec solution is fair and unbiased in its recommendations is important to the CEO. There can be dire consequences of losing track of the fairness goal to both the people being exposed to the AI solution and the organization, as in the case reported by the *Economist* in January 2021, where the Dutch Prime Minister had to resign because an algorithm designed to detect welfare fraud had wrongly accused more than 20,000 beneficiaries (Economist 2021).

Privacy awareness is necessary to avoid the illegal use of private data. Privacy is a human right and must be addressed properly by the organization (Stahl et al. 2022). In fact, the XAI techniques themselves may, in some cases, violate the privacy laws by revealing private data (Barredo Arrieta et al. 2020), for instance by crowdsourcing image classification (Mauri and Bozzon 2021). In the Pinot StudRec solution, if the AI group includes external data in the solution and combines data from different sources, information about privacy-sensitive areas may emerge, for example about political affiliation or sexual orientation, even though the applying student did not provide this information.

Causality relates to inferring causality from data. When ML models discover correlations, the humans involved must be able to use their in-depth domain knowledge to judge if there is also causality (Barredo Arrieta et al. 2020). In some cases, showing correlation may be enough. For instance, in the Pinot case, it helps the company to know that there is a correlation between the profiles of students that StudRec suggests they employ and previously selected students who

are now full-time employees and show exemplary performance. An assumption of causality may be supported by false correlations. Therefore, it is also important to be aware of the problem of spurious correlation in AI design, as one may find false correlations in massive datasets (Gandomi and Haider 2015).

Transferability means the possibility that the AI solution can be used for purposes other than what it is designed for. Such explainability means that humans can evaluate if the AI solution can actually be used with good results in a different setting. For instance, Pinot could try to use the StudRec solution for all its recruitment, but it may cause problems because the non-student applicants might be evaluated on more recruitment criteria than the students.

Accessibility is the ability of non-AI-experts to understand how the AI solution works. For example, if the Pinot recruiter does not understand what the AI solution does, there are two choices: upskill the recruiter or increase the explainability of the AI solution.

Confidence implies how stable the model is over time, that is, how confident one can be in the AI solutions' results. For example, if the Pinot StudRec solution selected different students every time it ran, it would evoke low confidence and therefore be unfit for recruitment.

Informativeness is the most frequent expectation of explainability and speaks to the information that is needed for the user to understand the decision generated by the model. As the problem that the user is trying to solve may differ from the problem that the AI solution is intended to solve, there may be a need for information to connect the two. The user needs enough information to take the correct decision from a different, possibly wider perspective than what the AI solution provides. In the StudRec solution, what kind of informativeness could be achieved depends on who will use it, and for what purpose. The recruiter may need informativeness to review the list of students deemed employable by StudRec and, if he/she is to be part of that process, in deciding who will actually get employed. Alternatively, StudRec may directly inform the student who provided their data if she/he got the job or not, and thereby the informativeness goal would be directed towards the student.

Interactivity relates to the AI solution's ability to interact with users to help them with their work. For example, depending on how the task structure is designed, the focus could be the Pinot recruiter's or the student's interaction with the AI solution.

Trustworthiness determines whether a user will act in accordance with an AI solution's decision or override it. In the Pinot case, showing trust would mean the recruiter accepts the list of students provided by the AI solution.

To achieve Barredo Arrieta et al.'s (2020) explainability goals, some of the AI solutions will have to provide XAI. XAI software has various explanatory techniques that can be used to increase explainability (Hall 2018). XAI can, for instance, be user-oriented visualizations, interfaces, and toolkits (Wang et al. 2019). Other factors than the XAI can also influence the explainability. This because, the quality of predictions from ML models depends on many aspects: the availability and quality of the training data, and what kind of training is possible (Jordan and Mitchell 2015), as well as defining the training goals used to evaluate the performance of the AI solution (Lebovitz et al. 2021). In addition, the choice of ML model can affect prediction quality. The design choice regarding the explainability of the AI solution can lead to a trade-off between the accuracy of the predictions and the level of explanation needed (Deeks 2019; Rai 2020), because an increased prediction accuracy can reduce the explainability of the AI solution. This is so because an ML model that optimizes for accuracy of prediction will generally be too complex for humans to understand with regards to how it reaches its results. In order for the ML model to be more explainable to humans, it will have to be simpler and, thus, result in less prediction accuracy. This means that the XAI level needed to support user trust in the AI solution's decisions may make the decisions more inaccurate and thereby less trustworthy. This underscores the importance of not pushing the AI solution towards too much explainability but rather focusing on reducing the need for explanations in the organizational design. The explainability goals are part of different basic activities performed in the organization, and how these activities are structured into tasks affects how the goals are achieved. Therefore, we propose to use sociotechnical systems design principles for designing the task structure in a way that supports achieving the explainability goals.

3 Sociotechnical systems design and IT architecture

Sociotechnical systems theory (STS) originates in Trist & Bamforth's (1951) study of new mechanical equipment in British coal mines, which demonstrated how the introduction of new technology hampered effective work organization. The introduction of new technology challenged the established social working system. These findings formed the basis for a design approach that recognizes integrated interactions between people and technology as a necessity. An important design principle in sociotechnical systems design (STSD), emphasized by Herbst (1974, 1993), is that decisions should be made at the lowest organizational level possible. In addition, when those closest to the technology are allowed to give input into the design of the system, the workers benefit from the challenge, variety, feedback, and teamwork that is involved in taking responsibility for the performance of the system (Pasmore et al. 2019). This requires

empowering employees to be able to cope with problems on their own by redesigning the way they work, increasing each employee's autonomy (Pasmore 2001). Such autonomy can be designed by following the design parameters and structures described by de Sitter (Achterbergh and Vriens 2011; de Sitter et al. 1997; Vriens and Achterbergh 2011). These STSD structures and principles are based on a functional model of the organization with four basic activities: *strategic regulation* activities that set and adapt the organization's goals, the *regulation by design* activities that improve the people, technology, and task structure of the organization, the *operational regulations* that handle the disturbances that occur, and the *primary processes* where the product or service is produced and supported (see Table 1).

According to Govers and Südmeier (2016), STSD principles are a toolkit for designing IT solutions that are non-bureaucratic. For instance, Govers and Van Amelsvoort (2018) describe how an IT architecture can follow the principle of parallelized value streams to adhere to organizational design parameters for flexible organizations, in what they call an archipelago architecture. The archipelago architecture presents different user interfaces for each of the parallel value streams that people use. In the organizational architecture description by Govers and Van Amelsvoort (2018), there seems to be an assumption that there are always people present in the primary processes. There are nodes (people/teams) that transform input into output based on information from the IT solutions. We interpret this to mean that the primary process is a 'moving target' in that the work performed by the people in the organization is always considered to be the primary process. This may lead to challenges when trying to describe the task structure of automated solutions. When the primary processes are fully automated, for instance in bank applications, the people involved are the design and development team, that is, they are performing regulation by design activities. We would therefore like to use the term primary process both for organizations where the primary processes are performed by people with the help of technology, and for organizations where the primary processes are performed by technology alone. What is described here is the difference between augmented and automated work. AI solutions that augment work are becoming abundant (Davenport and Miller 2022), for instance, radiologists studying mammography images with AI support to improve the chances of cancer detection (Rodríguez-Ruiz et al. 2019). Automated primary processes can be seen in work settings that are highly structured (Davenport and Miller 2022), for instance providing banking or insurance services to customers (Iansiti and Lakhani 2020). What complicates the organizational design process even further is that some AI solutions have a level of agency that make them able to perform autonomously (Kaplan and Haenlein 2020), although in quite limited environments (Autor 2015). The AI solution can also *self-automate* in that it can learn and adapt to perform the tasks previously performed by humans, in what is called Intelligent Automation (Coombs et al. 2020). This means that when redesigning the work system to include AI, it is both complex and crucial to understand what kind of automation degree is feasible and desirable, and what kind of augmentation to provide. STSD principles address the handling of uncertainty and variance and provide advice on what type of activities to automate and what activities to augment, which means that the necessary explainability becomes clearer.

Organizational design in STSD consists of two parts; first, the functional model of the basic activities in organizations mentioned above, and second, the structural design parameters to describe who does what. The three regulation activities, operational regulation, regulation by design, and strategic regulation, are called the control structure (Achterbergh and Vriens 2011). The use of the term 'regulation' is based on the idea that the organization has essential variables that it must keep within certain boundaries, for instance, that the revenue needs to be positive. If it is not, then one or more regulation activities are necessary to bring the essential variables back to within the defined boundaries. The control, therefore, both deals with disturbances that occur as well as with setting targets (Achterbergh and Vriens 2009; de Sitter et al. 1997).

The structural design parameters consist of three groups (Achterbergh and Vriens 2011; Vriens and Achterbergh 2011). The first group is linked to the operational tasks in the organization and how the production is structured to handle, for instance, order types. It breaks down into *degrees of*

Table 1 The basic activities of an organization and examples of activities

Type of structure	Basic activity	Examples of activities
Control structure	• Strategic regulation	• Set and adapt goals
	• Regulation by design	• Improve the way the organization works: the people, the technology, and the task structure
	• Operational regulation	• Monitor, assess, and act to handle disturbances
Production structure	• Primary process	• Prepare, produce, and support

functional concentration, *differentiation*, and *specialization*. The second group is linked to the control tasks, especially the design of the regulatory tasks. The regulatory activities may similarly be divided into *parts, aspects* and *specialization*. *Parts* are the differentiation of process steps (correlating to the operational/production tasks' *degree of differentiation*). *Aspects* are the differentiation of different kinds of regulation (correlating to the operational/production tasks' *degree of functional concentration*). And *specialization* are smaller sub-activities (correlating to the production tasks' *degree of specialization*). The third group concerns the relation between the primary processes and the regulatory activities, looking at the *degree of separation*. The values of the design parameters are defined so that high parameter values are fragmented designs, while low parameter values are a flow-oriented designs. To obtain an adaptive organization, the design parameters should have low values, because high parameter values make it harder to have an overview of the entire process, understand what is going on, and act on that.

3.1 First structural design parameter group: operational tasks and how the production is structured to handle order types

A high degree of functional concentration means that all order types are sent through the same task structure, and the people doing the activities are grouped together based on function. A low degree of functional concentration is parallelized into different order types. The parallelized order types can be divided by, for instance, type of product, type of client, or geography. In Fig. 1, we have illustrated the high functional concentration setting where all order types are handled by different functional departments, one after the other (fragmented). This as opposed to the low functional concentration setting where different order types are divided between multifunctional teams that, together, handle the whole work process (flow oriented). In the high concentration setting, this means that there may be a lot of variety in the orders. In addition, because they work sequentially, there is little learning and adaptation between the functional departments. In the low functional concentration setting, the orders are divided into order types to lower the variety of the orders. The lowered variety makes it easier for the team to handle all the variety it is exposed to. This leads to increased learning and adaptation.

3.2 Second structural design parameter group: control tasks

The second group of design parameters entails differentiation, that is *parts, aspects*, and *specialization* of regulatory tasks (strategic regulation, regulation by design, and operational regulation). The same principle applies here: wider responsibility, that is, a lower degree of differentiation and/or specialization, is better for the adaptiveness of the organization. This means that if a person is performing operational regulation tasks, there is more adaptiveness if one person performs monitoring, assessment and action compared to if these tasks are given to different people.

3.3 Third structural design parameter group: the relation between operational and regulatory activities

The third group of design parameters concerns the *degree of separation* between the primary processes and the control structure—the regulatory activities. This concerns whether the workers in the primary processes have the responsibility and possibility to handle the variety they encounter and improve their own work. A high degree of separation is seen when someone performs the primary processes, and someone else monitors the work and interferes when something needs to be corrected. A high degree of separation can also be seen when someone decides how to perform a primary process task, and someone else performs it. In a low degree of separation, the same person/team is expected to perform the primary process tasks, evaluate the quality of the work,

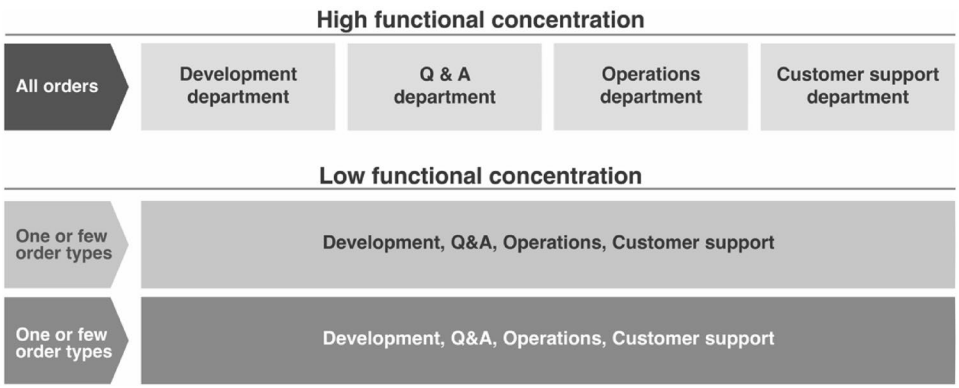

Fig. 1 High and low functional concentration exemplified

and correct it accordingly. In the latter case, the people are allowed to learn and improve their way of working.

4 The AI design and development team

In an AI solution design team, there will typically be competence on how to build ML models, for instance in the form of a data scientist (Davenport and Patil 2012). The explainability of the AI solution is considered the responsibility of the data scientist because they build the ML models (Minh et al. 2022). To make data available, there is often a need for a data engineer creating the data pipeline (Tamir et al. 2015). The AI solution may be placed in a cloud (for instance Microsoft Azure) or on local servers, and depending on this, there may be a need for cloud or local server operational experts. The team may be delivering directly to the market/environment, and, to handle external variety that occurs from conflicting and changing customer demands (Van Amelsvoort 2016), user experience designers are needed to find out about the users' needs and to design a user interface that enables interaction and informativeness. If the AI solution operates in a digital ecosystem, the AI solution may be installed into an automatic task structure which interacts with other IT solutions instead of people, for instance in the case of fraud detection software that monitors credit card transactions and trigger actions like disabling a credit card. In that case, it may be necessary for a programmer to integrate the systems. The way the team works is usually decided by the team, perhaps with the help of a facilitator or team leader, and it is beneficial to the teamwork if the team creates temporal alignment and learns about the effects of their work (Wulff and Finnestrand 2022). The team can be an integral part of the organization's work, or it may be put together for a specific project. The competence may come from inside the organization and/or be provided by consultants. When using consultants it is vital that internal people are also part of the team (Tabrizi et al. 2019), and that the team engages the organization in its work. The team can be a functional team with the focus on delivering a product, or it may be a sociotechnical team focused on changing the way the organization works (Achterbergh and Vriens 2019). Of course, both functional and sociotechnical teams do change how the organization works, but in the case of the sociotechnical team, the changes are reflected upon, and the work is done with a higher degree of involvement from the rest of the organization to enable motivation, adoption, and integration of the AI solution. Hence, our preference is for a sociotechnical team as the development team. This may also mean acquiring new capabilities, both regarding the possibilities and different ways of task structure redesign (Gong and Ribiere 2021). Also, according to Stahl et al. (2022), although the developers are not trained in ethics, they are expected to handle it. Hence, the competence of the people creating and using the AI solution may need to be increased. Such upskilling could be a part of the competences that the sociotechnical team provides when leading and facilitating organizational change and (re)designing task structures, and may be covered by roles like change manager, leader and/or organizational designer. The design team may also benefit from being diverse as that can make it more aware of bias issues (Daugherty et al. 2019).

The people who will use the AI solutions for augmented purposes will participate in designing and redesigning the AI solution with the sociotechnical team, because they know how to do the work and they will need to integrate the solution into their future work. This may also involve union representatives and leaders.

The data that the team uses can come from internal and external data. When using internal data, the data may come from IT solutions already used in the organization as well as from people in the organization recording the data. To ensure that the team receives the right quality of data, this may create team tasks like understanding what data is available from IT solutions, contacting IT solution providers to get access to application programming interfaces (APIs), extracting and/or saving data and perhaps triggering actions based on occurrences in the IT solution. All these data might be necessary to train an ML model and thereby affects the quality of the output from the AI solution and its explainability. The people who currently record or are expected to record data in future will benefit from being involved in the design process so that they understand why the data are useful and how they can learn from the data. For external data, there may often be defined APIs, that is, external databases made available via interfaces on the Internet.

Creating an AI development team and deciding on the other actors involved in making design decisions is, therefore, based on broader considerations than the team being proficient in using AI tools. To aid the work of the AI development team and its surrounding stakeholders, we demonstrate how the use of design parameters can affect the ML model's explainability.

5 Discussion

The need for re-examining the principles for designing flexible organizations has increased with the introduction of the AI technologies that are used today, especially ML. ML is able to handle more variety, doing non-routine work that increases the need for understanding the choices the AI solution makes, in real time or later. However, demanding that the ML explains itself can reduce the accuracy of its predictive power. Therefore, it is necessary to understand how STSD contributes to the design of well-functioning task

structures and how it contributes to changing or reducing the need for XAI.

5.1 The basic activities and the explainability goals

With regards to AI, we expect the strategic regulation on an organizational level to set ethical goals for the organization's behavior towards both its employees and its customers. This means that we expect the people involved in strategic regulation to show an interest in the *fairness* and *privacy awareness* goals. The act of performing strategic regulation for the organization may also be supported by AI (Keding 2021), although intuition still plays an important part (Liebowitz et al. 2019). In such cases, we would expect *causality* and *transferability* to be of interest.

Regulation by design activities include all the design and redesign of the AI solution and the task structure. This determines what the next version of the organization will look like regarding who performs what tasks and with what technology and level of autonomy. One of the use cases for explainability in regulation by design is to be able to correct and redesign the AI solution (Ammanath et al. 2020), that is, to improve its *accessibility*. To build *trustworthiness* in the AI solution, regulation by design activities should—by following the low degree of separation parameter value—include operational knowledge. For instance, the CEO in the Pinot case could include a recruiter in the AI solution design team, and representatives of the loan consultants could be included in the Strich et al. (2021) case. The *confidence* that the AI solution is providing as accurate decisions as necessary is something that can be checked over time, either by comparing results from two different version of an ML model, monitoring the decisions, or by being notified of errors.

If it is an augmented solution, the operational regulation and primary processes are activities where people are interacting with the AI solution. In that case, the explainability goals we will expect to find are *informativeness* of the AI solution to the people involved as well as useful *interactivity*. We believe that the *accessibility* of the AI solution will be important here as well and, as mentioned before, its *trustworthiness*.

For the stakeholders and designers of the AI solution, it will be beneficial to understand how the different explainability goals are experienced in different basic activities. The actual design of the AI solution performs with regards to two groups of explainability goals: First, those where a pattern emerges as the AI solution is run many times, which are *fairness*, *privacy awareness*, *causality*, *transferability*, and *confidence*. These patterns can/should emerge as part of regulation by design activities (see Table 2), and we suggest that these are named *pattern goals*. Second, those where each run of the AI solution creates an experience of achieving the goal or not, which are *accessibility*, *informativeness*, *interactivity*, and *trustworthiness*. These goals are experienced as part of the operational tasks of the primary processes and operational regulation, and we suggest that these are named *experience goals*. We recognize that fairness—or rather unfairness—may be experienced in the moment of, for instance, a loan approval rejection, but whether this actually is an instance of unfairness can only be seen over time. We believe that what we call 'sociotechnical team' is necessary to successfully handle all these design challenges. The following sociotechnical advice will, however, also be beneficial to other types of teams.

5.2 The sociotechnical systems design parameters and the explainability goals

The STSD parameters are divided into three groups: the design parameters for the production structure, for the control structure, and for the connection between the production and the control structure. We will first look into the explainability goals for the two first groups.

The required explainability can be affected by whether the workers in the production structure are in an organizational design with a high or low functional concentration. For instance, a design team develops an AI solution to find patterns in customer feedback to come up with a prediction about what the customer's next call will be about. The goal of the explainability will be the *informativeness* for the person answering the call. The workers in the customer support department in a high functional concentration setting, for example a separate customer support department which is not involved in the deliveries of the organization, need more information on what has happened so far to understand why the customer is calling. In comparison, when the customer support is an integral part of a low functional concentration team, the person doing the support knows what has happened to date and better understands the context of the customer and why he/she calls. If the AI solution, in addition to providing information, also suggests a response to the customer, its *trustworthiness* needs to be high.

As mentioned earlier, a task may be divided into smaller or larger parts depending on the variety of the internal or external environment. This *degree of differentiation* of operational tasks concerns whether the same or different people perform the different parts of the primary process (such as preparing material, producing, and providing support). The degree of differentiation of operational tasks is high if each task is performed by different people, and low if several tasks are performed by the same people. An example of a high degree of differentiation is the case of the loan consultants by Strich et al. (2021) where the task structure design is such that a loan consultant enters the client information into an AI solution, and then the AI solution produces an

Table 2 The division of XAI goals into two groups and sociotechnical advice on how to increase the chance of achieving the goals

Basic activity	XAI goals	Sociotechnical advice
Strategic regulation	• Ethical and accountability goals, both vis-á-vis employees and clients	• Reach agreement on how important it is to produce fair and ethical AI solutions.
Regulation by design	• Pattern goals: Fairness, Privacy awareness, Causality, Transferrability, Confidence	• Seek participation from operations to gain enough operational knowledge, and to bring AI design knowledge back to operations • Decide who must be involved to understand the consequences of getting it wrong • Choose a team that ensures bias awareness • Check whether the operational task structure has low degree values before redesigning the process, and check whether it will have low degree values after the redesign • Debate who to involve inside (and outside) the organization to quality control the data, to monitor how the model handles privacy, and how likely it is that the ML model can be transferred to other areas
Primary process and Operational regulation	• Experience goals: Accessibility, Informativeness, Interactivity, Trustworthiness	• Check if there are people performing tasks in the primary processes or operational regulation • If yes, ask what kind of main structure will be used: 1. people and AI performing both primary processes and operational regulations, or 2. people performing just one of them and the AI the other? • If option 2., ensure minimizing the problems that this division of tasks produces with regards to operational explainability goals

answer that the loan consultant passes on to the client. This happens without the employee having any overview of or say in the decision. If the client asks why they did not get the loan, the loan consultant is expected to provide answers that she/he does not have. Therefore, the goal of the explainability will be to increase its *informativeness* so that the loan consultant can provide satisfactory answers to the client. To provide such an XAI, there may be a need for choosing a simpler ML model than the optimal one, thereby increasing the risk of giving loans to the wrong people and/or wrongly declining loans.

The *degree of specialization* regards how specialized the work is. In a high degree of specialization scenario, the task is divided into different specializations, and this can be true for both control and production structure tasks. An AI solution designed for a primary process with a high degree of specialization would be, for instance, that one AI solution delivers classification with regards to mammography results while another delivers classification with regards to X-rays for orthopedic issues. This could fit well if the task structure is divided in the same way, that is, different radiologists evaluate the two, but if the same radiologist evaluates both results, they will have to learn to use two different AI solutions. The need for *informativeness* and *interactivity* might be reduced if the specialization of the AI solution matched how specialized the radiologists are. In addition, the explainability here is directed towards the consistency of the AI solution, and how much *confidence* one can have in its results. If the AI solution is to be used in a new setting, the *transferability* is assessed. For instance, can the AI solution for detecting breast cancer be used for other cancer forms as well.

The third group of design parameters considers the connection between the production structure and the control structure. In a high degree of separation structure, the production and the control are performed by different entities, while in a low degree of separation structure production and control are performed by the same entity. In a low degree of separation, one strives to design the technology so that either the people perform both the primary process and the operational regulation with help from the technology, or that the technology performs both types of activities. An example of the latter case would be a credit card system where the transactions are automatic, and the fraud detection (control) is done by AI which triggers locking of the account when fraud is detected (Roy et al. 2018).

As we see it, based on the literature, a high degree of separation structure where the primary process and operational regulations are performed by different entities has two main designs (Table 3). First, a design where people perform the primary processes and AI solutions perform the operational regulation, in what is called 'algorithmic management'. Second, a design where technology performs the primary processes and people do the operational regulation, that is, the human is expected to monitor and assess the technology, and act if it fails. We suggest calling this 'monitoring' for lack of a better word.

Both these high degree of separation designs, as presented in Table 3, are problematic from an organizational performance perspective. The first, algorithmic management, reduces employee control and thereby the possibility to improve the work (Kellogg et al. 2020). In algorithmic management, one can find AI solutions that perform, for instance, task allocation and scheduling (Schildt 2017), which we understand as operational regulation activities. Where technology performs the operational regulation on humans, the *informativeness* lies in conveying to the human what he/she is supposed to be doing. In other cases, the algorithmic management controls the workers and either interferes or notifies the leaders in case of errors (Herrera et al. 2018). The *interactivity* then can mean explaining to the human why the AI solution saw it necessary to interfere in the work process, for instance by stopping a production line to prevent a harm from happening. The detrimental effects on employees of algorithmic management designs can be mediated by transparency and worker control over the output (Parent-Rocheleau and Parker 2022), that is, that workers are allowed to do regulation by design activities. In the case of human monitoring, that is, where technology performs the primary processes and people perform the operational regulation, the XAI must compensate for the lack of situational awareness (Endsley 1995; Endsley and Kaber 1999) that follows when the human is not performing the primary process, in what is often called humans-out-of-the-loop problems (Banks et al. 2014). Such a lack of situational awareness increases the need for the *informativeness* and *interactivity* of the AI solution. A possible way to reduce the problems of this design is to ensure that the people performing the operational regulation participate in designing the technology, which means that they perform regulation by design activities. This increases the possibility that they will understand the explanations from the XAI and its *accessibility.*

One reason for choosing the monitoring task structure design is because the consequences of something going wrong in the primary process would be severe, which influences the explainability. The need for explainability increases with the severity of the consequences of an unfair or faulty automatized decision (Wang et al. 2019). If a book recommendation system fails to provide us with interesting book recommendations, the consequences are low (Doshi-Velez and Kim 2018). In contrast, an autonomous ship that fails may set in motion a high-consequence situation (Utne et al. 2017). In such situations the expectations put on the XAI may be unrealistic because, when the task structure design is problematic, the necessary explainability may be impossible to provide. The European Commission has deemed decisions that affect people's lives or careers to be of high consequence (Cappelli et al. 2020). This includes obtaining and using personal data (Kochan 2021). To ensure that such decisions are understandable, the General Data Protection Regulation (GDPR) includes a rule on automated decision-making (European Commission 2018) to ensure that decisions which are legally binding or affect people's lives significantly are given with explanations (Goodman and Flaxman 2017). Likewise, banking is regulated to ensure that the use of AI solutions is ethical and accountable (Maree et al. 2020). Such challenges place even greater demands on the designers of the algorithms to avoid creating unfair decision systems when automating AI (Shrestha et al. 2019), to understand the consequences for the work and organizing (Faraj et al. 2018), and to aim for an optimal task structure.

The successful introduction of ML models into organizations will in part depend on the perceived value of the classifications, predictions, or prescriptions that the AI solution delivers. This means that it is important to be able to choose the best ML model for the job and to optimize the task structure that the AI solution will function in. If organizations are going to allow AI solutions to interact with us as customers and users, it is a necessity to achieve some or all of the explainability goals (Barredo Arrieta et al. 2020). We have described how the design choices regarding the task structure the AI solution slots into influence its explainability. Therefore, if an organization wanting to benefit from

Table 3 High degree of separation operational design structures

Basic activities / Type of design	Algorithmic management	Monitoring
Operational regulation	• Technology incl. AI	• People
Primary process	• People	• Technology incl. AI

the use of AI evaluates and redesigns the organization to the appropriate degree values, it increases the probability that its investment will not be wasted.

6 Conclusion

In this paper, we have examined explainability from a wider perspective than XAI design alone. The choices made on which data to train the model with, how to handle biases, who to involve in the design, and how to design the task structure can reduce or change the need for XAI. As we have demonstrated, the required XAI changes depending on the organizational design principles used. The goal is for a high-quality organizational design with optimal explainability, an organization that increasingly includes autonomous technological systems and has enough flexibility to handle the disturbances that occur by using low value flow designs. This will prevent organizations from introducing AI solutions into low-quality organizational designs, and then demanding that the XAI fix a situation that the fragmented design has caused to start with.

There are two organizational challenges we would like to contribute to with this paper. First, that the AI solution actually reaches a stage where it produces value, and second, that the AI solution is integrated into the organization in a way that creates meaningful jobs for the people involved and supports the productivity of the organization. Our contribution was inspired by real-life business problems and was encouraged by researchers calling for more organizational research (f.i. Raj and Seamans 2019; von Krogh 2018). We were unable to find research that provided design principles for organizational design with AI. Thus, we concluded that a theoretical paper laying a foundation that could afterwards be tested by empirical research would be beneficial to the field of AI research. In order to study such a topic, it is appropriate to use a systemic theory aimed at recognizing and redesigning the task structure. Much of the existing AI research seems to be oriented towards the decision-making in a company (f.i. Delen and Ram 2018; Shrestha et al. 2019). Amongst organizational design theories, we chose de Sitter's STSD because it provides a scale for each parameter value going from high to low, which can be helpful for describing how the organization works today and how it can work in future. In contrast, the Cherns (1976, 1987) principles can be seen as more normative, although, also the de Sitter STSD promotes organizations with low parameter values. According to Mohr and Van Amelsvoort (2016), STSD is the least-known approach to organizational improvement. Nevertheless, we believe that STSD provides theory and design parameters that are necessary for today's organizations. We would like to encourage further research into organizations' readiness for applying AI solutions into their task structures to find out how and if the introduction of AI solutions includes organizational design evaluations and changes.

The are several implications of arguments in this paper. First, designing an AI solution is part of an organizational design change and as such will benefit from understanding and implementing organizational design parameters. Second, when designing a task structure where AI will interact with workers in performing operational tasks, the need for explainability varies with how the task structure is designed. Third, the *pattern* goals and the *experience* goals of explainability are reached in different ways. The *experience* goals may be achieved by redesigning the task structure. The achievement of the *pattern* goals will emerge over time as the AI solution is used, but can and should be addressed in the initial design to strive for goal fulfillment. To understand whether the goals are achieved, the AI solution's performance must be evaluated and redesigned accordingly. We have described the difference between a functional development team and a sociotechnical one, and hope to have illustrated the benefits of taking a systemic approach and engaging a sociotechnical team. Based on the design reflections we have presented here, we would encourage the design of XAI to be supported by organizational design knowledge.

Acknowledgements This research is supported by The Research Council of Norway under Grant number 290629 in an Industrial PhD project for the software company Kantega AS. We would like to thank our colleague Dr. Jon Espen Ingvaldsen for insightful and inspiring help in revising this paper. A warm thank you to the reviewers for insightful comments that pinpointed improvements to be made.

Author contributions KW: conceptualization, investigation, writing, editing. HF: structuring, advising, reviewing, and editing.

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital).

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare that they have no financial benefit from the direct application of our research.

References

Achterbergh J, Vriens D (2009) Organizations: social systems conducting experiments. Springer, Berlin and Heidelberg. https://doi.org/10.1007/978-3-642-00110-9

Achterbergh J, Vriens D (2011) Cybernetically sound organizational structures II: Relating de Sitter's design theory to Beer's viable system model. Kybernetes 40(3–4):425–438. https://doi.org/10.1108/03684921111133665

Achterbergh J, Vriens D (2019) Organizational development: designing episodic interventions. Routledge, London and New York. https://doi.org/10.4324/9781315695228

Ammanath B, Hupfer S, Jarvis D (2020) Thriving in the era of pervasive AI. Deloitte's State of AI in the Enterprise. Deloitte Insights, Deloitte AI Institute, Colombus, Ohio. https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/about-deloitte/deloitte-cn-dtt-thriving-in-the-era-of-persuasive-ai-en-200819.pdf. Accessed 25 Nov 2022

Autor DH (2015) Why are there still so many jobs? The history and future of workplace automation. J Econ Perspect 29(3):3–30. https://doi.org/10.1257/jep.29.3.3

Babüroğlu ON, Selsky JW (2021) Toward reconfiguring sociotechnical systems design: digitally infused work systems and the "Platform-STS." In: Shani AB, Noumair DA (eds) Research in organizational change and development, vol 29. Emerald Publishing Limited, Bingley, pp 63–87. https://doi.org/10.1108/S0897-301620210000029004

Banks VA, Stanton NA, Harvey C (2014) Sub-systems on the road to vehicle automation: hands and feet free but not 'mind' free driving. Saf Sci 62:505–514

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58(June):82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barro S, Davenport TH (2019) People and machines: partners in innovation. MIT Sloan Manag Rev 60(4):22–28

Benbya H, Davenport TH, Pachidi S (2020) Special issue editorial: Artificial Intelligence in organizations: current state and future opportunities. MIS Q. Executive 19(4):ix–xxi. https://doi.org/10.2139/ssrn.3741983

Biran O, Cotton C (2017) Explanation and justification in machine learning: a survey. IJCAI-17 workshop on explainable AI (XAI), Melbourne, Australia, 20 August. http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf Accessed 16 May 2022

Buhrmester V, Münch D, Arens M (2021) Analysis of explainers of black box deep neural networks for computer vision: a survey. Mach Learn Knowl Extr 3(4):966–989

Bussone A, Stumpf S, Sullivan DO (2015) The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. 2015 International Conference on Healthcare Informatics, 21–23 October. https://openaccess.city.ac.uk/id/eprint/13150/1/ICHI_2015_CameraReady.pdf. Accessed 16 May 2022

Cappelli P, Tambe P, Yakubovich V (2020) Can data science change human resources? In: Canals J, Heukamp F (eds) The future of management in an AI world: redefining purpose and strategy in the fourth industrial revolution. Springer International Publishing, Cham, Switzerland, pp 93–115. https://doi.org/10.1007/978-3-030-20680-2_5

Castelvecchi D (2016) Can we open the black box of AI? Nature News 538(7623):20–23. https://doi.org/10.1038/538020a

Cherns, A (1976)The Principles of Sociotechnical Design. Hum Relat 29(8):783–792

Cherns, A (1987) Principles of Sociotechnical Design Revisited. Hum Relat 40 (3):153–162.https://doi.org/10.1177/001872678704000303

Colaner N (2022) Is explainable artificial intelligence intrinsically valuable? AI Soc 37(1):231–238. https://doi.org/10.1007/s00146-021-01184-2

Coombs C, Hislop D, Taneva SK, Barnard S (2020) The strategic impacts of Intelligent Automation for knowledge and service work: an interdisciplinary review. J Strateg Inf Syst 29(4):101600. https://doi.org/10.1016/j.jsis.2020.101600

d'Alessandro B, O'Neil C, LaGatta T (2017) Conscientious classification: a data scientist's guide to discrimination-aware classification. Big Data 5(2):120–134. https://doi.org/10.1089/big.2016.0048

Daugherty PR, Wilson HJ, Chowdhury R (2019) Using Artificial Intelligence to promote diversity. MIT Sloan Manag Rev Digital 60(2)

Davenport TH, Miller SM (2022) Working with AI: real stories of human-machine collaboration. MIT Press, Cambridge

Davenport TH, Patil D (2012) Data scientist: the sexiest job of the 21st century. Harv Bus Rev 90(5):70–76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

de Sitter LU, Hertog JFd, Dankbaar B (1997) From complex organizations with simple jobs to simple organizations with complex jobs. Hum Relat 50(5):497–534. https://doi.org/10.1177/001872679705000503

Deeks A (2019) The judicial demand for explainable Artificial Intelligence. Columbia Law Rev 119(7):1829–1850

Delen D, Ram S (2018) Research challenges and opportunities in business analytics. Journal of Business Analytics 1(1):2–12. https://doi.org/10.1080/2573234X.2018.1507324

Doshi-Velez F, Kim B (2018) Considerations for evaluation and generalization in interpretable machine learning. In: Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, Van Gerven M, van Lier R (eds) Explainable and interpretable models in computer vision and machine learning. Springer, Cham, pp 3–17. https://doi.org/10.1007/978-3-319-98131-4_1

Economist T (2021) The fraud that wasn't. The Economist 438(9229). https://www.economist.com/europe/2021/01/23/a-benefits-scandal-sinks-the-dutch-government. Accessed 16 May 2022

Eiband M, Schneider H, Bilandzic M, Fazekas-Con J, Haug M, Hussmann H. (2018). Bringing Transparency Design into Practice 23rd International Conference on Intelligent User Interfaces, Tokyo, Japan, 7–11 March. https://doi.org/10.1145/3172944.3172961

Emery F, Trist EL (1965) The causal texture of organizational environments. Hum Relat 18:21–32

Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. Hum Factors 37(1):32–64. https://doi.org/10.1518/001872095779049543

Endsley MR, Kaber DB (1999) Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42(3):462–492. https://doi.org/10.1080/001401399185595

Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, Van Gerven M, van Lier R (2018) Explainable and interpretable models in computer vision and machine learning. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-98131-4

European Commission (2018) Are there restrictions on the use of automated decision-making? Directorate-General for Communication. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/dealing-citizens/are-there-restrictions-use-automated-decision-making_en. Accessed 16 May 2022

Faraj S, Pachidi S, Sayegh K (2018) Working and organizing in the age of the learning algorithm. Inf Organ 28(1):62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manag 35(2):137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gong C, Ribiere V (2021) Developing a unified definition of digital transformation. Technovation 102:102217. https://doi.org/10.1016/j.technovation.2020.102217

Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision-making and a "right to explanation." AI Mag 38(3):50–57. https://doi.org/10.1609/aimag.v38i3.2741

Govers M, Südmeier P (2016) Applying enterprise information technology from a socio-technical perspective. In: Co-creating humane and innovative organizations: evolutions in the practice of socio-technical system design. Global STS-D Network, Portland ME, pp. 289–302

Govers M, Van Amelsvoort P (2018) A socio-technical perspective on the design of IT architectures: the lowlands lens. Manag Stud 6(3):177–187. https://doi.org/10.17265/2328-2185/2018.03.003

Govers M, Van Amelsvoort P (2019) A socio-technical perspective on the digital era: the lowlands view. Eur J Workplace Innov 4(2):142–159. https://doi.org/10.46364/ejwi.v4i2.589

Guest D, Knox A, Warhurst C (2022) Humanizing work in the digital age: Lessons from socio-technical systems and quality of working life initiatives. Hum Relat 75(8):1461–1482. https://doi.org/10.1177/00187267221092674

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A Survey of methods for explaining black box models. ACM Comput Surv 51(5):1–42. https://doi.org/10.1145/3236009

Gunning D (2016) Explainable artificial intelligence (xai) IJCAI 2016 Workshop on Deep Learning for Artificial Intelligence, New York, NY (USA), 10 July

Hall P (2018) On the art and science of machine learning explanations KDD '19 XAI Workshop, Anchorage, AK, 4–8 August. https://arxiv.org/pdf/1810.02909.pdf

Heinrichs B (2022) Discrimination in the age of artificial intelligence. AI Soc 37(1):143–154. https://doi.org/10.1007/s00146-021-01192-2

Herbst DPG (1974) Designing with minimal critical specifications. In: Herbst PG (ed) Socio-technical design: strategies in multidisciplinary research. Tavistock Publications, London, pp 294–302

Herbst DPG (1993) A learning organization in practice, M/S Balao. In: Trist E, Murray H (eds) The social engagement of social science: a tavistock anthology, vol II. University of Pennsylvania Press, Philadelphia, pp 409–416

Herrera JLL, Figueroa HVR, Ramírez EJR (2018) Deep fraud. A fraud intention recognition framework in public transport context using a deep-learning approach. In: 2018 international conference on electronics, communications and computers (CONIELECOMP), 21–23 Feb 2018

Iansiti M, Lakhani KR (2020) Competing in the age of AI, strategy and leadership when algorithms and networks run the world. Harvard Business Review Press, Boston

Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260. https://doi.org/10.1126/science.aaa8415

Kaplan A, Haenlein M (2020) Rulers of the world, unite! The challenges and opportunities of artificial intelligence. Bus Horiz 63(1):37–50. https://doi.org/10.1016/j.bushor.2019.09.003

Keding C (2021) Understanding the interplay of artificial intelligence and strategic management: four decades of research in review. Management Review Quarterly 71(1):91–134. https://doi.org/10.1007/s11301-020-00181-x

Kellogg KC, Valentine MA, Christin A (2020) Algorithms at work: the new contested terrain of control. Acad Manag Ann 14(1):366–410

Kim B, Doshi-Velez F (2021) Machine learning techniques for accountability. AI Mag 42(1):47–52. https://ojs.aaai.org/index.php/aimagazine/article/view/7481. Accessed 16 May 2022

Kochan T (2021) Artificial intelligence and the future of work: a proactive strategy. AI Mag 42(1):16–24. https://ojs.aaai.org/index.php/aimagazine/article/view/7387. Accessed 21 Oct 2021

Lawless WF, Mittu R, Sofge DH, L, (2019) Artificial intelligence, autonomy, and human-machine teams—interdependence, context, and explainable AI. AI Mag 40(3):5–13. https://doi.org/10.1609/aimag.v40i3.2866

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. Manag Inf Syst Q 45(3b):1501–1525. https://ssrn.com/abstract=3839601

Leslie D (2019) Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute, London. https://doi.org/10.5281/zenodo.3240529

Liang Y, Li S, Yan C, Li M, Jiang C (2021) Explaining the black-box model: a survey of local interpretation methods for deep neural networks. Neurocomputing 419:168–182. https://doi.org/10.1016/j.neucom.2020.08.011

Liao QV, Gruen D, Miller S (2020). Questioning the AI: informing design practices for explainable AI user experiences proceedings of the 2020 CHI conference on human factors in computing systems, Honolulu. https://doi.org/10.1145/3313831.3376590

Liebowitz J, Chan Y, Jenkin T, Spicker D, Paliszkiewicz J, Babiloni F (2019) If numbers could "feel": How well do executives trust their intuition? VINE J Inf Knowl Manag Syst 49(4):531–545. https://doi.org/10.1108/VJIKMS-12-2018-0129

Maree C, Modal JE, Omlin CW (2020) Towards responsible AI for financial transactions. 2020 IEEE symposium series on computational intelligence (SSCI), Canberra, Australia, 1–4 December

Mauri A, Bozzon A (2021) Towards a human in the loop approach to preserve privacy in images. In: CEUR workshop proceedings. http://ceur-ws.org/Vol-2947/paper6.pdf. Accessed 23 May 2022

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

Minh D, Wang HX, Li YF, Nguyen TN (2022) Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55(5):3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mohr BJ, Van Amelsvoort P (2016) Co-Creating Humane and Innovative Organizations, vol 1. Global STS-D Network, Portland ME

Parent-Rocheleau X, Parker SK (2022) Algorithms as work designers: how algorithmic management influences the design of jobs. Hum Resour Manag Rev. https://doi.org/10.1016/j.hrmr.2021.100838

Pasmore W (2001) Action Research in the Workplace: the socio-technical perspective. In: Reason P, Bradbury H (eds) Handbook of action research: participative inquiry and practice Sage, London, pp 38–47

Pasmore W, Winby S, Mohrman SA, Vanasse R (2019) Reflections: sociotechnical systems design and organization change. J Chang Manag 19(2):67–85. https://doi.org/10.1080/14697017.2018.1553761

Rai A (2020) Explainable AI: from black box to glass box. J Acad Mark Sci 48(1):137–141. https://doi.org/10.1007/s11747-019-00710-5

Raj M, Seamans R (2019) Primer on artificial intelligence and robotics. J Org Design 8(1):1–14. https://doi.org/10.1186/s41469-019-0050-0

Rodríguez-Ruiz A, Krupinski E, Mordang J-J, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, Mann RM (2019) Detection of breast cancer with mammography: effect of an Artificial Intelligence support system. Radiology 290(2):305–314. https://doi.org/10.1148/radiol.2018181371

Rossi F (2019) AI ethics for enterprise AI. Harvard. https://economics.harvard.edu/files/economics/files/rossi-francesca_4-22-19_ai-ethics-for-enterprise-ai_ec3118-hbs.pdf. Accessed 16 May 2022

Roy A, Sun J, Mahoney R, Alonzi L, Adams S, Beling P (2018) Deep learning detecting fraud in credit card transactions. In: 2018 systems and information engineering design symposium (SIEDS)

Russell S, Norvig P (2021) Artificial Intelligence a modern approach, 4th edn. Pearson, Hoboken

Schildt H (2017) Big data and organizational design—the brave new world of algorithmic management and computer augmented transparency. Innovation 19(1):23–30. https://doi.org/10.1080/14479338.2016.1252043

Shao Y, Cheng Y, Shah RU, Weir CR, Bray BE, Zeng-Treitler Q (2021) Shedding light on the black box: explaining deep neural network prediction of clinical outcomes. J Med Syst 45(1):5. https://doi.org/10.1007/s10916-020-01701-8

Shrestha YR, Ben-Menahem SM, von Krogh G (2019) Organizational decision-making structures in the age of Artificial Intelligence. Calif Manag Rev 61(4):66–83. https://doi.org/10.1177/0008125619862257

Sørmo F, Cassens J, Aamodt A (2005) Explanation in case-based reasoning-perspectives and goals. Artif Intell Rev 24(2):109–143. https://doi.org/10.1007/s10462-005-4607-7

Stahl BC, Antoniou J, Ryan M, Macnish K, Jiya T (2022) Organisational responses to the ethical issues of artificial intelligence. AI Soc 37(1):23–37. https://doi.org/10.1007/s00146-021-01148-6

Strich F, Mayer A-S, Fiedler M (2021) What do i do in a world of Artificial Intelligence? Investigating the impact of substitutive decision-making AI systems on employees' professional role identity. J Assoc Inf 22(2):304–324. https://doi.org/10.17705/1jais.00663

Tabrizi BN, Lam E, Girard K, Irvin V (2019) Digital transformation is not about technology. Harv Bus Rev. https://hbr.org/2019/03/digital-transformation-is-not-about-technology Accessed 24 Nov 2022

Tamir M, Miller S, Gagliardi A (2015) The data engineer. Available at SSRN 2762013. https://doi.org/10.2139/ssrn.2762013

Trist, EL, Bamforth, KW (1951) Some Social and Psychological consequences of the Longwall Method of Coal- Getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. Hum Rel 4(1):3–38. https://doi.org/10.1177/001872675100400101

Utne IB, Sørensen AJ, Schjølberg I (2017) Risk management of autonomous marine systems and operations. International Conference on Offshore Mechanics and Arctic Engineering

Van Amelsvoort P (2016) Human talent mobilization: improving both quality of working life and productivity by organizational design in the lowlands. In: Mohr BJ, Van Amelsvoort P (eds) Co-creating humane and innovative organizations, vol 1. Global STS-D Network, Portland ME, pp 73–98

Van Lent M, Fisher W, Mancuso M (2004) An explainable artificial intelligence system for small-unit tactical behavior. In: IAAI'04: Proceedings of the 16th conference on Innovative applications of artificial intelligence, San Jose, California, 25–29 July. https://dl.acm.org/doi/abs/10.5555/1597321.1597342. Accessed 16 May 2022

von Krogh G (2018) Artificial intelligence in organizations: new opportunities for phenomenon-based theorizing. Academy of Management Discoveries. https://doi.org/10.5465/amd.2018.0084

Vriens D, Achterbergh J (2011) Cybernetically sound organizational structures I: de Sitter's design theory. Kybernetes 40(3):405–424. https://doi.org/10.1108/03684921111133656

Wang D, Yang Q, Abdul A, Lim BY (2019) Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI conference on human factors in computing systems, Glasgow, 4–9 May

Wolf CT (2019) Explainability scenarios: towards scenario-based XAI design Proceedings of the 24th International Conference on Intelligent User Interfaces, Marina del Ray, California. https://doi.org/10.1145/3301275.3302317

Worren N (2018) Organization design: simplifying complex systems. Routledge, London and New York

Wulff K, Finnestrand H (2022) It is like taking a ball for a walk: on boundary work in software development. AI Society 37:711–724. https://doi.org/10.1007/s00146-021-01175-3