Privacy Challenges for Process Mining in Human-centered Industrial Environments

Felix Mannhard, Sobah Abbas Petersen and Manuel Fradinho Oliveira

Citation:
Felix Mannhard, Sobah Abbas Petersen and Manuel Fradinho Oliveira(2018) Privacy Challenges for Process Mining in Human-centered Industrial Environments. In: 2018 14th International Conference on Intelligent Environments (IE), Rome, Italy, 25-28 June 2018
DOI: 10.1109/IE.2018.00017

Privacy Challenges for Process Mining in Human-centered Industrial Environments

Felix Mannhardt*, Sobah Abbas Petersen* and Manuel Fradinho Oliveira*
*SINTEF Technology and Society
Trondheim, Norway
{felix.mannhardt, sobah.petersen, manuel.oliveira} @sintef.no

***Abstract*—Operators in industrial manufacturing environments are under pressure to cope with increasing flexibility and complexity of work. The automation of manufacturing requires operators to adopt new techniques and shifts the focus from low-complexity repetitive tasks to dealing with the execution of high-complexity tasks in cooperation with machines. The emergence of wearable technologies makes it possible to equip operators with miniaturized sensors that may be used to determine the physical and mental stress experienced by operators. Process mining technologies are suited to analyze such sensor data in the context of the manufacturing process with the ultimate goal of improving the operator's well-being through re-organization of work and the work place. However, the storage and processing of such highly personalized data comes with many privacy challenges. Whereas there are many potential benefits, such as improve the work environment, there are also many justified reasons for operators to oppose the processing of their data. Apart from employee concerns, data protection regulations, such as EU GDPR (Europe's General Data Protection Regulation), imposes many compliance challenges for the design of a process mining systems dealing with personal data. We contribute an analysis of the privacy challenges of using process mining on data recorded from sensorized operators in human-centered industrial environments. Guided by privacy research and the regulation imposed by the GDPR, we describe guidelines for privacy in process mining systems.**

I. Introduction

The well-being of operators in industrial manufacturing environments is crucial to the success of organizations. Operators experience physical and mental stress due to increasing flexibility and complexity of work. Stressful situations and bad work practices should be detected and mitigated. Re-organization of the work processes or changes to the work place can be possible mitigation strategies. With the emergence of wearable technology [1], it is increasingly possible to equip the operator with sensors that may be used to detect such stressful situations. Moreover, sensors are ubiquitous in the industrial environment to steer the actual operations. Through trends like Industry 4.0 machines and sensors are increasingly connected and execution data is being stored [2]. This enables to capture the execution of activities on the shop-floor in the form of event sequences and correlated physiological sensor data from the operator. Such captured event data can be used by process mining technology [3], to provide an accurate view on what really happens on the shop-floor. This makes it possible to put detected stress situations in the context of work processes. Insights obtained can, then, be used to improve the work situation based on evidence.

However, a substantial obstacle in the acceptance of such data collection and processing are *privacy concerns* both among employees [4] and data protection regulations imposed by governments, such as the recently introduced EU GDPR (Europe's General Data Protection Regulation) [5]. Privacy concerns have been raised since personal information was stored in databases and could be processed using computers [6]. However, since it is possible — through advances in the amount of storage and processing power available — to store and process virtually all information that might be of interest (Big Data), the right to privacy has been in the focus of public attention. The seemingly never ending collection of data by large corporations such as Google and Facebook has raised public awareness on privacy questions [7]. Therefore, when introducing process mining into human-centered industrial environments privacy should be considered as *first-class citizen*: Privacy should be introduced by design and not as an afterthought.

Whereas there has been plenty of research on what constitutes privacy [8]–[10] and it its role in information systems engineering [11]–[14], there is a clear *gap in the research on privacy in the field of process mining*. Since process mining analyses are often more interested in the organizational processes rather than individual people, personal data is not necessarily processed. However, when events include information about employees or customers, then privacy challenges appear. Specifically, if process mining is used to improve work processes with a focus on the well-being of operators, privacy challenges need to be addressed.

This paper contributes an analysis of the privacy challenges encountered when employing process mining in human-centered industrial environments. We introduce a concrete application scenario in which process mining is used to help analyzing the well-being of human operators. In the application scenario both data about the work executed and physiological signals of operators is collected (e.g., using data recorded in an intelligent environment). Guided by this scenario and regulations like the GDPR, we identify technological and organizational challenges for the application of process mining. Then, we describe a set of preliminary guidelines that are applicable to application scenarios in which process mining uses personal data and personalized support is provided.

The paper is structured as follows. Section II introduces background on privacy and process mining. Section III describes an application scenario of process mining and identifies privacy challenges. Section IV presents a set of guidelines and Section V concludes the paper and sketches future work.

II. Background

We introduce background on privacy, process mining, and briefly review the related literature. First, we describe the basics of the right to privacy in the context of collection and analysis of (personal) data. Then, we look at the literature on process mining in industrial environments and existing work on privacy considerations in process mining analyses.

A. Privacy and GDPR

Privacy is generally considered the fundamental human right to be let alone and free from interruptions or intrusions. Information privacy is the right to have some control over how your personal information is collected and used. The focus on privacy has gained increasing attention over recent years due to the vast amounts of data that is constantly gathered by the systems that we use and how this data is used to provide services for us by various systems and service providers. The recent GDPR [5] has introduced changes to the privacy and data protection regulations which can have significant implications in the way we design our systems and treat the data that are used for providing services to our users. Many systems capture and use data pertaining to individual humans (*data subjects*) or personal data, with the intent of providing personalized services (e.g., the recommendations that pop up on web browsers).

The GDPR defines personal data as "any information relating to an identified or identifiable natural person (*data subject*)". The requirements for anonymization or pseudonymization are enhanced and require that personal data is processed with the aim to irreversibly prevent the identification of the individual to whom the data relates to. The GDPR provides new rights to the data subjects where they now have control of their data, improving data transparency and empowerment of data subjects.

The GDPR is focused on the protection of personal data, not merely the privacy of personal data. The protection of personal data and the privacy of personal data are not the same; data protection is about securing the data against unauthorized access while privacy of data is about authorized access, such as who has it and who defines it. Data protection is a technical issue while data privacy is a legal issue. The distinctions between the two concepts are important to understanding how one complements the other. It is important to understand that data protection is essential to ensure data privacy; i.e. if someone has unauthorized access to the data, then privacy cannot be guaranteed. For example, when someone uses a credit card in a shop, she trusts the shop and the payment system to protect their data from unauthorized access (e.g. criminals). At the same time, she trusts the shop to honor data privacy by not misusing the information even though they have access to it. Data privacy, thus, goes beyond technological solutions into the softer aspects of an organization such as trust in the organization.

The GDPR requires *privacy-by-design*, which calls for the inclusion of data protection from the onset of the designing of systems, rather than an addition. This implies considering privacy of users and data protection to ensure privacy and trust as a part of the design, along with other design aspects such as the functionality, performance and user interface. It also emphasizes the importance of considering it throughout the life-cycle of the system from conception to the design, development, deployment and maintenance. In summary the GDPR introduces the following new elements.

- *Location* — GDPR applies to all organizations who controls or processes personal data of data subjects residing in the EU, regardless of their location.
- *Penalties* — Under GDPR organizations in breach of GDPR can be fined up to 4% of annual global turnover or 20 Million (whichever is greater).
- *Consent* — Consent must be clear and distinguishable from other matters and provided in an intelligible and easily accessible form, using clear and plain language. It must be as easy to withdraw consent as it is to give it.
- *Breach* — Notification is mandatory, where a data breach is likely to result in a risk for the rights and freedoms of individual.
- *Data Controller* and *Data Processor* — A controller is the entity that determines the purposes, conditions and means of the processing of personal data, while the processor is an entity which processes personal data on behalf of the controller. Data controllers and processors must involve the Data Protection Officer (DPO) as relevant.
- *Data Protection Officer* — Organizations have a bigger responsibility to assess the impacts of the privacy implications of processing personal data. The role of a DPO is not mandatory except when the data processing operations require regular and systematic monitoring of data subjects on a large scale, or when special categories of data or data relating to criminal convictions and offenses are processed. The main tasks of the DPO are to inform and advise the data processors and controllers and to monitor compliance with the GDPR. [15]
- GDPR gives several new rights to the data subjects:
  - *right to access* — to obtain from the data controller confirmation as to whether or not personal data concerning them is being processed, where and for what purpose. Further, the controller shall provide a copy of the personal data, free of charge, in an electronic format.
  - *right to portability* — to receive the personal data concerning them, which they have previously provided in a 'commonly use and machine readable format' and have the right to transmit that data to another controller.

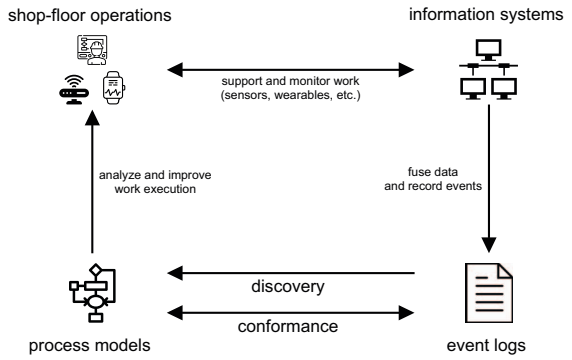

Fig. 1. Process mining in human-centered industrial environments is used to leverage the data captured of shop-floor operations through sensors and applications for the analysis and improvement of the work execution. Data needs to be fused into events that relate to specific work processes and can, therefore, be analyzed in the context of discovered or existing process models. Adapted from an overview in [3].

  - *right to oppose processing* — to cease further dissemination of the data, and potentially have third parties halt processing of the data, such as profiling and automated decisions. The data subject has the right to stop the secondary use of the data.
  - *rights to be forgotten* — to have the data controller erase his/her personal data. The condition for erasure is that the data no longer is relevant to the original purposes for processing, or a data subject withdraws consent.
- *Privacy-by-design* — which calls for the inclusion of data protection from the onset of the designing of systems, rather than an addition.

Clearly, the privacy right associated with the GDPR and the requirement of privacy-by-design has implications for any kind of data capturing and data analysis. We focus on the implications for the application of process mining methods that process data originating from the shop-floor in human-centered industrial environments in both real-time or non-real time. Next, we briefly introduce process mining and the considered types of analysis.

B. Process Mining in Industrial Environments

With growing computing power and storage capacity of today's IT systems, organizations can *store information about all their activities* that are supported by IT systems or that are observed using sensors. Typically, the execution of a case results in a sequence of events being recorded. In general, such an execution trace, also denoted log trace, contains at least: the timestamps of activity executions and names or identifiers of the executed activity. Each *log trace* groups together the activities performed in one instance of a recurring process. An event log containing a set of several such log traces captures a data-driven view on the process execution. Thus, process mining adds the notion of process instances (also denoted cases) and activity sequences in comparison to other data analytics methods, which operate mainly on flat data files. As depicted in Figure 1, process mining uses event logs to *analyze the actual execution of processes* [3]. The two main types of process mining are: *process discovery* and *conformance checking*. *Process discovery* aims to automatically discover accurate process models from event logs, e.g., to detect how work is actually performed and to reveal workarounds. *Conformance checking* aims to compare the real process execution with existing de-jure models, e.g., to pinpoint deviations and project information about the execution onto a process model.

In the context of human-centered industrial environments, several processes are worth analyzing using process mining. Typical candidates for activities are, e.g., the individual assembly tasks performed by operators and logistical activities around the supply with parts and materials. However, data sources are not limited to the execution of activities. It is possible to fuse sensor data from wearable device and machines (e.g., in an intelligent environment) together with the execution of work activities and overlay sensor data with a discovered process model. Furthermore, automatic reasoning techniques can be applied to detect undesired situations such as mental or physical stress encountered by operators. Such events can also be used as input for process mining. For example, in [16], [17] the behavior of workers is used to determine their workload.

Despite this opportunity, compared to other fields only little research has been conducted on the application of process mining in industrial environments such as manufacturing. In [18] a Supervisory Control and Data Acquisition (SCADA) system is analyzed based on artificial data with the goal of explaining production shut-downs. Similarly, in [19] the general applicability of process mining in logistics is shown, also by using artificial data. In [20], several process discovery algorithms are compared on event logs from a coffee machine manufacturing company. Unfortunately, no process models are shown. In [21] a process discovered from data about media manufacturing is presented. Finally, the work in [22] is an example that contextual factors can also be used. In this case the manufacturing cost was predicted using a method based on process mining. Thus, the application of process mining in industrial environments is promising, yet more research is required.

C. Privacy in Process Mining

As motivated earlier, one very important aspect in the application of process mining, is privacy. Already in 2012, the Process Mining Manifesto [23] stated that event logs of the highest quality should address privacy and security considerations adequately and the development of privacy-preserving process mining methods would be required for cross-organizational mining [23]. Liu et al. consider the privacy of internal event logs and process models in a cross-organizational process discovery setting in [24]. A few very valuable guidelines for the privacy-aware application of process mining, from a practical viewpoint and mainly in a consulting context, are given by Rozinat et al. in [25].

Thus, very little research has been conducted on the privacy challenges faced by organizations that want to employ process

mining methods in a privacy-aware manner. Moreover, in the light of the GDPR regulations and its increasing relevance, privacy cannot be seen merely as a problem of event log data quality, in the context of a cross-organizational setting, or as a problem for external consultants. If personal data is involved in the analysis, which is very likely to be the case if individual operators can be identified, additional organizational processes and technological solutions around the processing of the data are required. In the areas of Big data and data mining the privacy challenge has been long recognized [26]. A large stream of computer science research is concerned with the development of privacy-preserving algorithms [27], methods like differentiable privacy [28] and homomorphic encryption are possible algorithmic solutions. Another related research perspective is how to use result from data analytics responsibly. In [29] an overview and practical guidelines are given. There has also been research on general engineering strategies to build information systems that are aware of privacy challenges both on a technological and an organizational level [11], [12], [14].

One could argue that privacy considerations in process mining are similar to those in the works just described. However, the specific input to process mining, event logs, and the fact that we consider the personal data of operators in an industrial environment warrants a discussion that is specifically aimed at process mining in our scenario. Using both highly sparse, sequential log data with timestamped events and sensor data from wearable sensors is challenging from a privacy viewpoint.

III. Privacy Challenges for Process Mining in Human-centered Industrial Environments

There are many benefits that may be reaped by applying process mining methods to the collected data derived from personal data. For example, improvements to the work environment and the way of working may reduce the physical and mental stress of operators. However, the storage and processing of such personalized data comes with many privacy challenges. There may be many justified reasons for operators to oppose the processing of their data. Operators may fear that collected data may compromise their career progression or may not be suitably protected against access from adversaries. As previously described, regulations like the GDPR put the data subjects (here the operators) in control of their data. This leads to both technological and organizational challenges for using captured data containing personal data that can be associated to a particular individual.

A. Prototypical Application Scenario

We identify privacy challenges for process mining in a human-centered industrial environment and illustrate them based on the data flow in the prototypical application scenario of process mining that is shown in Figure 2. In this scenario, we consider a process mining system[1] in which data flows through six distinct stages: data source, data capture, primary use, data storage, secondary (re)use, and data removal. Thus, we analyse the flow of the data from the time it is captured at its source to its removal from the system.

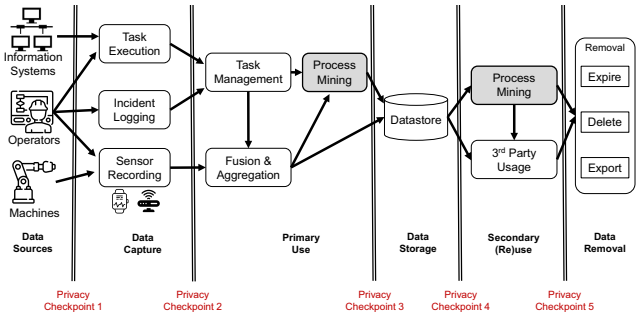

Fig. 2. A typical application scenario of process mining in human-centered industrial environments. Data flows through six distinct stages from its inception at a data source to its removal. Process mining can be applied both for primary use (e.g., to support the task execution) or for secondary use of the data (e.g., to diagnose inefficiencies causing stress). Data passes several privacy checkpoints for which we identify privacy-related guidelines.

1) Data Source: In the application scenario, we envision three main sources for data that is used for process mining: data coming from *machines*, data originating from *operators*, and data cross-referenced from *information systems*. Clearly, data that originates directly from operators and their actions may be categorized as personal data.

2) Data Capture: Data is captured by sensors from both machines (e.g., operational parameters) and operators (e.g., wearables or manual registration), by systems supporting the task execution (e.g., manufacturing execution systems), and by systems logging incidents. Sensors record physiological parameters of operators (e.g., heart-rate, blood pressure, movement, etc.) or operational parameters of machines.

3) Primary Use: The primary use of the captured data – besides for operational purposes such as task management – in our application scenario is the application of real-time process mining methods (online process mining) that directly support the operator's work. To apply process mining methods, we need to combine data coming from sensors together with data from the task management (i.e., execution and incident events). Moreover, some of the low-level sensor data needs to be aggregated and abstracted to form higher-level events that can be used in process mining. For example, high-frequency data from multiple sensors can be combined with other contextual data (e.g., task execution) to assess whether the operator is stressed. Machine learning methods or complex event processing engines may be used to this end.

In this stage, mainly methods for predictive process mining that give operational support to the operator are used. Several *predictive process mining* that could be applied have been proposed: queue delay prediction [30], remaining processing or service time prediction [31], [32], prediction of the most likely next activities [33], and compliance prediction [34], [35]. In the context of a human-centered industrial environment, these

[1]The same applies to data used by a consultant in an ad-hoc manner.

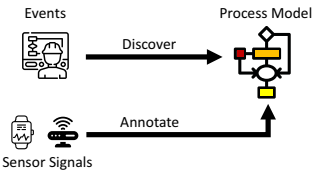

Fig. 3. A possible application of process mining is the discovery of a process model from the sequence of events generated by the work of the operator. The model is annotated information on the physical and mental stress of operators.

techniques may be useful for numerous reasons. First, they may help to prevent incidents related to work safety (e.g., skipping of safety checks). Second, they might reduce stress due to overworking (e.g., early prediction that performance goals cannot be met). Third, they can help to organize the work in better ways (e.g., by pro-actively monitoring predicted stress levels).

4) Data Storage: If used for more than its primary purpose the personal data needs to be stored. Storage may be provided through a database or in the form of event log files. The duration of storage depends on the envisioned secondary use of the data.

5) Secondary (Re)use: At this stage, the personal data is used in ways that do not affect the operator in the short-term and, thereby, it is difficult for the operator to realize the full implications at the time of consent. As an example, one may determine key characteristics to form a worker profile to be used in future recruitment of workforce, but this same worker profile can determine whom should be phased out of the existing workforce. When using process mining technology, there are two options: (1) to consider a product- or product-line-centric view on the data and discover end-to-end process models that include all activities conducted to produce an single product; or (2) to consider all the activities executed by a single operator on a specific day as a process instance. The first view could reveal undocumented workarounds and re-work. The second view would reveal how the work of operators is organized. When enriching the event log with personal data (e.g., the stress level of the individual operator) it is possible to annotate discovered models with aggregated indicators for physical and mental stress of the operator (Figure 3), which reveals the "pain points" in the work. This allows to use data-aware discovery techniques such as [36], [37] to reveal stressful situations in the context of the process. Moreover, conformance checking methods [38] could be used to check for deviations between existing work documentation and the real execution. This could reveal gaps in the work documentation or compliance issues.

6) Data Removal: Data is permanently removed from the data storage and not available for analysis anymore. Aggregated data, e.g., indicating the most frequent process variants and mean processing times and stress levels may be retained for comparison. However, the raw data is deleted.

This application scenario exemplified the prototypical flow of the data, i.e., the transitions in the life-cycle of data and possible ways to use data for process mining. Based on it we now identify technological and organizational challenges that organizations face when attempting to ensure privacy while still being able to conduct process mining towards meeting business goals of the organization.

B. Technological Privacy Challenges

We identified the following technological challenges when applying process mining to data captured in the chosen application scenario. Technological challenges have an impact on how to achieve *privacy-by-design* or *privacy-by-architecture* [11], [12]. Note that these challenges need to be addressed by all the systems in our application scenario that handle data for transfer, storage and processing [11].

1) Minimization Challenge: Data minimization is a core principle of any approach to ensure privacy. For example, the GDPR states as principle that data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimisation)" [5]. Here, we refer to the technological challenge of how to minimize the storage and processing of personal data while still being able to draw conclusions. In the process mining application scenario, personal data may be processed on two levels: (1) the sensor signal level, which captures physiological signals of operators and (2) the timestamp and work sequence level, which captures the timing and ordering of the work done by operators. Whereas the privacy issue regarding physiological data is clear, the problem of capturing timing and sequence of executed work may be less obvious. From such data it may not be possible to directly identify a single operator. However, for complex processes the sequence of activities is very sparse[2]. Such sparse data — which is the core data source for process mining — is at risk of re-identification by using limited knowledge about the operator even if personal identifiers have been removed [39], [40]. The challenge is how to limit the possibilities of such re-identification while still being able to use it for process mining.

2) Aggregation Challenge: One option to address the minimization challenge would be to avoid processing data at the level of individual operators. After all, the goal of applying process mining is to increase the well-being of operators through improvement of the overall process. However, aggregating data may remove the possibility to give personalized support to individual operators in a real-time fashion. Thus, a possibility would be to limit the individual data processing to the primary use stage and only retain aggregated data thereafter. The challenge is how to aggregate event data such that privacy is guaranteed. We are not aware of research in the process mining field that aims to tackle this question. However, concepts like k-anonymity could be employed and have been research in the context of sequence mining [41], which faces a similar problem.

[2]Already a process with 10 concurrent activities can be executed in $10! = 3,628,800$ different sequences [3].

3) Traceability Challenge: Regulations like GDPR require that data subjects need to provide consent for processing their data (*right to consent*). Moreover, they should also be able to retrieve their personal data (*right to access*) and remove it on request (*right to be forgotten*). The right to access is aimed more towards data of end-user of services such as provided by Google and Facebook. Nevertheless, it is important for organizations to be able to trace data through its life-cycle from the point it is captured until removal from their systems. A by-product of being able to trace back the origins of data that is used in a process mining analysis, is that it can build trust in the findings [25]. Since data is often fragmented across a wide variety of systems [42], providing traceability is an important challenge.

4) Monitoring and Transparency Challenge: Closely related to the traceability challenge, but from a different angle, is the challenge to monitor the actual usage of data. There need to be suitable technological solutions that monitor how stored data is used and whether its usage still complies with the initial purpose and the consent given by data subjects. In the process mining context this is particularly challenging since it is often used as exploratory tool for which the exact analysis goals are not pre-determined [43]. In contrast to the automatic processing of data for a pre-defined goal, e.g., classifying abnormal situations, which run largely automatically after an initial design phase, both design and execution of a process mining analysis are largely manual activities. Technological solutions would need to track who did what with the data and prevent data from escaping the monitoring capabilities of the system. For example, the often-available export functionality would defeat the purpose of such monitoring system and require other organizational measures.

5) Deletion Challenge: An important underlying premise of the GDPR is to give full control of the personal data to the owner, which in our case would be the operator. This means that the right to delete needs to be supported, but such action is non-trivial due to the ramifications of the impact that depends on the processing stage of the personal data. In the capture stage, the removal will limit the primary use and thereby reduce the benefits that the operator would get. The impact is also reduced to ensuring a satisfying user experience so the operator is aware of the implications. Should the personal data be used for its primary use, then the impact is considerable depending on how the data was processed and whether aggregation operations were carried out. The challenge needs to take into account the organizational processes and agreements with the operator as the deletion may not be possible in many circumstances where successful rollback is unfeasible, so it is crucial to ensure that an acceptable level of anonymity is achieved by the end of the primary use. However, even if it is possible to delete after processing, one is faced with organizational challenges where deletion may lead to a risk in operations (e.g., in the context of health and safety).

C. Organisational Privacy Challenges

Not all aspects of privacy can be ensured by technology alone. We identified the following five organizational challenges for using process mining.

1) Consent and Purpose Challenge: A central concept of the GDPR is the *consent* that needs to be asked for to process personal data. Consent needs to be retrieved before processing data and the option to withdraw consent must be readily available at any time. Very related to retrieving consent for the processing of personal data, is the challenge of defining the *purpose* of the collected data. When asking for consent, the purpose of processing needs to be clear to the operator. Clearly defining the purpose of data collection is specifically challenging for process mining since it is often conducted in a explorative fashion [43] in which multiple analysis iterations and multiple data collection rounds are needed to provide real value. Moreover, the GDPR requires that the purpose needs to be presented in a form understandable for the operator. A challenge is how to present the implications of sharing their data for process mining in a legible form.

2) Trust and Acceptance Challenge: The adoption of GDPR will bring to the forefront the trust relationship between the operator and the organization. The real issue is not the capture and processing of personal data, but understanding the purpose of the primary and secondary use of the data, and most importantly whether such usage may be detrimental to the interests of the operator (e.g., their career opportunities are stunted or more dramatically, the loss of employment). Addressing successfully the consent and purpose challenge will be an important step towards building trust in the organization, but ultimately, the operator needs to believe they are in control of their personal data and this requires transparency and accountability, form both the organization and the operator. However, some organizational processes raise technological challenges as giving full control of the personal data implies the ability of deleting it at any point in time or determining who has access to it.

3) Privacy vs. Benefits Challenge: The easiest option to ensure privacy is to avoid the collation of personal data altogether; however, without the necessary data, one cannot apply process mining with a particular purpose. Therefore, within the industrial context that we are addressing of an operator working on the shop-floor, there will always be a trade-off between privacy and benefits of sharing personal data. Naturally, the trade-off is not one-sided, but rather consists of a negotiation between an organization and the operators with regards to what benefits to consider. In any case, ensuring that operators understand the trade-off and have a voice in the negotiation is an important step towards addressing the trust and acceptance challenge.

4) Auditing Challenge: GDPR requires organizations to provide options for auditing of their processes that are related to the processing of personal data. So, it needs to be clear which process mining activities have been conducted and the full process from data capture to data removal has to be

transparent. Successfully addressing this challenge contributes to the increase of trust and acceptance by the operators.

5) Privacy Breaches Challenge: There are two dimensions to this challenge: *detection* and *mitigation*. Detection requires organizations to have the necessary mechanisms to recognize when security breach takes place and to quickly characterize precisely how it happened along with the impact assessment of the breach may cause. The successful analysis relies heavily on how the organization addresses the traceability, monitoring and auditing challenges. The other dimension of the challenge to consider is mitigation — what to do once a breach has taken place. Addressing how to prevent the breach may have wide implications that require restructuring the processes and systems in place, which in turn imply an investment cost that may be considerable. However, more importantly from the privacy perspective, is the speed and manner of disclosure with regards to the breach, taking into consideration that the perception of how it is handled my deteriorate or increase trust and acceptance by the operators. To compound the difficulty of how to address the challenge, one needs to acknowledge that cyber security in the manufacturing domain remains in its infancy.

We identified ten privacy challenges for process mining on personal data of operators. Whereas we believe to have covered the most important challenges in the described application scenario, this list should be seen as an initial proposal. Clearly, there are further challenges and their importance depends on the concrete scenario in which process mining is applied.

IV. Guidelines for Privacy-aware Process Mining based on the GDPR

We present a set of privacy guidelines in Table I for each of the privacy checkpoints that were identified in the application scenario for process mining (Figure 1). The guidelines are based on requirements imposed of the GDPR, work on privacy design strategies [12] and privacy design patterns [14] from software engineering as well as the requirements described in [13]. The guidelines may serve as a basis for the design of a system that incorporates process mining. We acknowledge that these guidelines are merely a starting point for actual design requirements and we do not claim that Table I is complete. Some of the guidelines may be in conflict with each other, e.g., the recommendation to anonymize data may be in conflict with the traceability.

TABLE I
Privacy guidelines for using process mining in human-centered industrial environments.

Data Capture	Inform operator (data subject) which data is collected. Inform about the duration of storage. Inform about the possibility for data removal and withdrawal of consent. Inform how data from several source is combined. Obtain consent from the operator for collecting data. Provide for privacy control and traceability of data.
Primary Use	Inform which real-time analysis will be conducted. Inform clearly about the risks and benefits of the envisioned analysis to build trust. Inform that the service cannot be provided without access to the data. Provide an option for the operator to delete data, at any point during or after the service. Aggregate the data to a suitable level of abstraction directly after usage.
Data Storage	Inform what the data is used for beyond the primary, real time use for the specific service(s). Obtain consent from the operator for secondary use of the data. Provide the option to determine how long the data can be stored for. Provide the option to determine who has access to the data that is stored and obtain consent as necessary. Anonymize unnecessary personal data with a suitable method before storing it. Encrypt the data while it is stored. Consider putting the operator in control of the encryption (client-side encryption).
Secondary Use	Inform which secondary service the data will be used for. Inform if the data is aggregated with other data. Inform if the data is reused in a manner that is not informed to the user previously. Inform if the data is exported or shared with a 3rd party. Obtain consent from the operator for secondary use of the data. Provide option to the operator to review the results obtained.
Removal	Provide an option to delete data, at any point during, at the end of or after a service (from storage). When data is deleted, ensure all information pertaining to the data are deleted, i.e., consider how to deal with analysis results obtained etc.

V. Conclusion

We introduced privacy and the GDPR in the context of process mining for human-centered industrial environments. Based on a concrete application scenario (Figure 1), we identified five technological and five organizational privacy challenges for the design of process mining systems and their application in practice. Furthermore, we presented several privacy guidelines for each of the privacy checkpoints of the assumed system that should be followed in its design and usage. This paper is a first proposal of privacy challenges for a concrete process mining application scenario. As future work, several directions are worthwhile pursuing: (1) the guidelines should be refined in a concrete process mining context and their usefulness be evaluated by testing the envisioned process mining system in real factories; (2) the challenges and guidelines could be generalized towards other application scenarios; and (3) the impact of designing a system in a privacy-aware manner on the acceptance among operators should be researched.

Acknowledgment

This research has received funding from the European Unions H2020 research and innovation programme under grant agreement no. 723737 (HUMAN).

References

[1] W. Barfield, *Fundamentals of Wearable Computers and Augmented Reality*. Taylor & Francis Ltd, 2017.

[2] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Business & Information Systems Engineering*, vol. 6, no. 4, pp. 239–242, Jun. 2014.

[3] W. M. P. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016.

[4] R. E. Silverman, "Bosses tap outside firms to predict which workers might get sick," 2016, wall Street Journal. [Online]. Available: http://www.wsj.com/articles/bosses-harness-big-data-to-predict-which-workers-might-get-sick-1455664940

[5] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," *Official Journal of the European Union*, vol. L119, pp. 1–88, 2016.

[6] A. R. Miller, "Personal privacy in the computer age: The challenge of a new technology in an information-oriented society," *Michigan Law Review*, vol. 67, no. 6, p. 1089, Apr. 1969.

[7] F. Stutzman, R. Gross, and A. Acquisti, "Silent listeners: The evolution of privacy and disclosure on facebook," *Journal of privacy and confidentiality*, vol. 4, no. 2, p. 2, 2013.

[8] Bélanger and Crossler, "Privacy in the digital age: A review of information privacy research in information systems," *MIS Quarterly*, vol. 35, no. 4, p. 1017, 2011.

[9] Smith, Dinev, and Xu, "Information privacy research: An interdisciplinary review," *MIS Quarterly*, vol. 35, no. 4, p. 989, 2011.

[10] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, Jan. 2015.

[11] S. Spiekermann and L. Cranor, "Engineering privacy," *IEEE Transactions on Software Engineering*, vol. 35, no. 1, pp. 67–82, Jan. 2009.

[12] J.-H. Hoepman, "Privacy design strategies," in *ICT Systems Security and Privacy Protection*. Springer, 2014, pp. 446–459.

[13] "Privacy and data protection by design - from policy to engineering," European Union Agency for Network and Information Security, Tech. Rep., 2014.

[14] M. Colesky, J. C. Caiza, J. M. D. Alamo, J.-H. Hoepman, and Y.-S. Martn, "A system of privacy patterns for user control," in *SAC 2018*. ACM Press, 2018, p. 2018.

[15] Baker & McKenzie LLP. (2016) Eu data protection officer - must have, nice to have or safe to ignore? [Online]. Available: http://globalitc.bakermckenzie.com/files/Uploads/Documents/Global ITC/13 Game Changers/BM-EU Data Protection Officer - Must Have, Nice to Have or Safe to Ignore.pdf

[16] J. Nakatumba and W. M. P. van der Aalst, "Analyzing resource behavior using process mining," in *BPM 2009 Workshops*. Springer, 2010, pp. 69–80.

[17] M. Park, M. Song, T. H. Baek, S. Son, S. J. Ha, and S. W. Cho, "Workload and delay analysis in manufacturing process using process mining," in *AP-BPM*, ser. LNBIP, vol. 219. Springer, 2015, pp. 138–151.

[18] B. Feau, C. Schaller, and M. Moliner, "A method to build a production process model prior to a process mining approach," in *INTELLI 2016*, 2016, pp. 143–146.

[19] T. Becker, M. Ltjen, and R. Porzel, "Process maintenance of heterogeneous logistic systems — a process mining approach," in *Dynamics in Logistics*, ser. LNLO. Springer, Sep. 2016, pp. 77–86.

[20] A. Bettacchi, A. Polzonetti, and B. Re, "Understanding production chain business process using process mining: A case study in the manufacturing scenario," in *CAiSE 2016*, ser. LNBIP. Springer, 2016, vol. 249, pp. 193–203.

[21] V. Muthusamy, A. Slominski, V. Ishakian, R. Khalaf, J. Reason, and S. Rozsnyai, "Lessons learned using a process mining approach to analyze events from distributed applications," in *DEBS 2016*. ACM Press, 2016.

[22] T. B. H. Tu and M. Song, "Analysis and prediction cost of manufacturing process based on process mining," in *ICIMSA 2016*. IEEE, May 2016.

[23] W. M. P. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. van den Brand, R. Brandtjen, J. Buijs, A. Burattin, J. Carmona, M. Castellanos, J. Claes, J. Cook, N. Costantini, F. Curbera, E. Damiani, M. de Leoni, P. Delias, B. F. van Dongen, M. Dumas, S. Dustdar, D. Fahland, D. R. Ferreira, W. Gaaloul, F. van Geffen, S. Goel, C. Günther, A. Guzzo, P. Harmon, A. ter Hofstede, J. Hoogland, J. E. Ingvaldsen, K. Kato, R. Kuhn, A. Kumar, M. La Rosa, F. Maggi, D. Malerba, R. S. Mans, A. Manuel, M. McCreesh, P. Mello, J. Mendling, M. Montali, H. R. Motahari-Nezhad, M. zur Muehlen, J. Munoz-Gama, L. Pontieri, J. Ribeiro, A. Rozinat, H. Seguel Pérez, R. Seguel Pérez, M. Sepúlveda, J. Sinur, P. Soffer, M. Song, A. Sperduti, G. Stilo, C. Stoel, K. Swenson, M. Talamo, W. Tan, C. Turner, J. Vanthienen, G. Varvaressos, E. Verbeek, M. Verdonk, R. Vigo, J. Wang, B. Weber, M. Weidlich, T. Weijters, L. Wen, M. Westergaard, and M. Wynn, "Process mining manifesto," in *BPM 2011 Workshops*, F. Daniel, K. Barkaoui, and S. Dustdar, Eds. Springer, 2012, pp. 169–194.

[24] C. Liu, H. Duan, Q. ZENG, M. Zhou, F. Lu, and J. Cheng, "Towards comprehensive support for privacy preservation cross-organization business process mining," *IEEE Transactions on Services Computing*, pp. 1–1, 2016.

[25] A. Rozinat and C. W. Gnther, "Privacy, security and ethics in process mining," Fluxicon, Tech. Rep., 2016. [Online]. Available: http://coda.fluxicon.com/assets/downloads/Articles/PMNews/Privacy-Security-and-Ethics-In-Process-Mining.pdf

[26] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, Jan. 2015.

[27] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *PODS 2001*. ACM Press, 2001.

[28] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[29] R. Clarke, "Guidelines for the responsible application of data analytics," *Computer Law & Security Review*, Dec. 2017.

[30] A. Senderovich, M. Weidlich, A. Gal, and A. Mandelbaum, "Queue mining for delay prediction in multi-class service processes," *Inf Syst*, vol. 53, pp. 278–295, 2015.

[31] W. van der Aalst, M. Schonenberg, and M. Song, "Time prediction based on process mining," *Inf Syst*, vol. 36, no. 2, pp. 450–475, 2011.

[32] A. Senderovich, C. D. Francescomarino, C. Ghidini, K. Jorbina, and F. M. Maggi, "Intra and inter-case features in predictive process monitoring: A tale of two dimensions," in *BPM*, ser. LNCS, vol. 10445. Springer, 2017, pp. 306–323.

[33] N. Tax, I. Verenich, M. L. Rosa, and M. Dumas, "Predictive business process monitoring with LSTM neural networks," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 10253. Springer, 2017, pp. 477–492.

[34] F. M. Maggi, C. D. Francescomarino, M. Dumas, and C. Ghidini, "Predictive monitoring of business processes," in *CAiSE*, ser. Lecture Notes in Computer Science, vol. 8484. Springer, 2014, pp. 457–472.

[35] A. Metzger, P. Leitner, D. Ivanovic, E. Schmieders, R. Franklin, M. Carro, S. Dustdar, and K. Pohl, "Comparing and combining predictive business process monitoring techniques," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 2, pp. 276–290, Feb. 2015.

[36] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, "Data-driven process discovery - revealing conditional infrequent behavior from event logs," in *CAiSE 2017*, ser. LNCS, vol. 10253, 2017, pp. 545–560.

[37] S. Schönig, C. Di Ciccio, F. M. Maggi, and J. Mendling, "Discovery of multi-perspective declarative process models," in *ICSOC 2016*, ser. LNCS, vol. 9936. Springer, 2016, pp. 87–103.

[38] F. Mannhardt, M. de Leoni, H. A. Reijers, and W. M. P. van der Aalst, "Balanced multi-perspective checking of process conformance," *Computing*, vol. 98, no. 4, pp. 407–437, 2016.

[39] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, May 2008.

[40] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–5805, Mar. 2013.

[41] A. Monreale, D. Pedreschi, R. G. Pensa, and F. Pinelli, "Anonymity preserving sequential pattern mining," *Artificial Intelligence and Law*, vol. 22, no. 2, pp. 141–173, Feb. 2014.

[42] H. van der Aa, H. Leopold, F. Mannhardt, and H. A. Reijers, "On the fragmentation of process information: Challenges, solutions, and outlook," in *CAISE 2015 Workshops*, ser. LNBIP. Springer, 2015, vol. 214, pp. 3–18.

[43] M. L. van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, "PM^2: A process mining project methodology," in *CAiSE*, ser. LNCS, vol. 9097. Springer, 2015, pp. 297–313.