
INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Luca
Nannini

PhD Thesis

# Explainability in Process Mining: A Framework for Improved Decision-Making

Santiago de Compostela, 2024

**Doctoral Programme in Information Technology Research**

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

TESE DE DOUTORAMENTO

# EXPLAINABILITY IN PROCESS MINING: A FRAMEWORK FOR IMPROVED DECISION-MAKING

Directores: Senén Barro Ameneiro, Alejandro Català Bolos
Tutor: Senén Barro Ameneiro

Luca Nannini

**PROGRAMA DE DOUTORAMENTO EN INVESTIGACIÓN EN TECNOLOXÍAS DA INFORMACIÓN**

SANTIAGO DE COMPOSTELA

**Ad Alberto Tartarini**

*a vój aŋdàr a ka...*

*«The highest form and final act of a semiotic system's structural organization is when it describes itself. This is the stage when grammars are written, customs and laws codified. When this happens, the system gains the advantage of greater structural organization, but loses its inner reserves of indeterminacy which provide it with flexibility, heightened capacity for information and the potential for dynamic development. The stage of self-description is a necessary response to the threat of too much diversity within the semiosphere*: *the system might lose its unity and definition, and disintegrate. Whether we have in mind language, politics or culture, the mechanism is the same*: *one part of the semiosphere* [...] *in the process of self-description creates its own grammar. This self-description may be real or ideal depending on whether its inner orientation is towards the present or towards the future.* [...] *The world-picture created in this way will be perceived by its contemporaries as reality. Indeed, it will be their reality to the extent that they have accepted the laws of that semiotics.* [...] *But the relationship of this metalevel of the semiosphere to the real picture of its semiotic "map" on the one hand, and to the everyday reality of life on the other, will be complex».*

J.Lotman - Universe of the Mind: A Semiotic Theory of Culture

## Agradecimientos

Contribución de ITN NL4XAI (Natural Language for Explainable AI) – financiado por el programa Horizon 2020 de la UE, beca Marie Skłodowska-Curie No 860621. Mi más sincero agradecimiento a mi supervisor principal, el Prof. Asociado Senén Barro Ameneiro, por guiarme sabiamente a través de este viaje de doctorado y a mi co-supervisor, el Prof. Asistente Alejandro Català Bolos, cuya paciencia y atención fueron fundamentales para cimentar mi investigación. Gracias al Prof. Asociado Manuel Lama Penin por sus comentarios sobre Minería de Procesos, y al Prof. Asociado Jose Maria Alonso por compartir su pasión por la IA Explicable. Al tratarse de un doctorado industrial, estoy agradecido a las personas de Minsait by Indra Sistemas. En primer lugar, gracias a Hitoshi Yano por proporcionarme una supervisión diaria y meticulosa. Gracias a los equipos de Onesait Language y Phygital Data. Un agradecimiento especial a Juan Prieto Vivanco por su orientación y a mi jefa Leticia Gómez Rivero y nuestro equipo de gobernanza de la IA. Por fin, a Raquel González Ibañez por su apoyo con los procedimientos administrativos. Agradezco a la red NL4XAI, particularmente a Raluca Tomoni Silva por asegurar el progreso del proyecto, y al Prof. Pablo Gamallo por las interesantes charlas. Gracias a quienes me acogieron durante las estancias: el grupo HMI de la UTwente, incluyendo a la Dra. Mariët Theune y Sumit Srivasatava; y el grupo WIS del Prof. Nava Tintarev y Alisa Rieger en la TU Delft. A nivel personal: primero a Daniele Mazzarisi, casi un hermano en medio del caos de Madrid; a Yoel Vincent Raphael, por su gran empatía y resolución para mejorar en la vida; a Lorenzo Vaquero, Pedro Gamallo, Yago Fontenla, Marcos Pichel, Marko Jaklin, Daniel Lesta y Peyman Fayyaz por los interminables días pasados en el CiTIUS; a Micaela Marini Higgs por haber sido lo mejor de Madrid para mí; a Sara Dominguez Rodriguez por su amor y entusiasmo. Un agradecimiento a Mirella, Chiara, Silvia y Enzo: a los nuevos capítulos que traen miedo pero oportunidades para redescubrir el cierre, incluso si físicamente están lejos. Finalmente, mi agradecimiento y admiración van para el pueblo de Ucrania, donde mi tío abuelo Giorgio Tartarini vivió en Ivano-Frankivsk hasta que la guerra de 2022 se lo llevó. A pesar de una historia de opresión, me enseñan que la democracia es un ejercicio continuo que vale la pena defender. No solo su futuro está en la Unión Europea, sino que el propio futuro de la UE se encuentra en esta nación: *Slava Ukraini*.

Julio 2024

# Resumo

Nunha era onde os coñecementos baseados en datos están a dar forma ao futuro das organizacións, a Minaría de Procesos (PM) xurdiu como un elemento transformador, ofrecendo oportunidades sen precedentes para analizar e optimizar procesos complexos. Con todo, o pleno potencial da PM segue sen aproveitarse debido aos retos persistentes na súa comprensión e na explicabilidade da tecnoloxía de procesos implementada, como os sistemas de IA. Esta tese establece unha investigación estrutural para explorar o papel da IA Explicable (XAI) para superar estas barreiras, promover unha maior adopción e compromiso de solucións de PM e, en última instancia, salvar a fenda entre as tecnoloxías avanzadas e a comprensión humana.

## Declaración de problema

A rápida proliferación de tecnoloxías de intelixencia artificial (IA) como a aprendizaxe automática en ámbitos críticos suscitou preocupacións sobre a súa transparencia e responsabilidade. En particular, os sistemas complexos de IA empregados en procesos operacionais de consecuencias representan ameazas derivadas da súa opacidade, o que aumenta os riscos de erros, prexuízos e danos sen vía de recurso. Isto subliña a necesidade crucial de técnicas XAI que fagan que estes modelos de IA sexan transparentes e interpretables.

Con todo, persisten deficiencias significativas en canto á aplicación rigurosa de solucións XAI na práctica do mundo real. Mentres que os algoritmos técnicos XAI por si mesmos poden alegar proporcionar transparencia, as restricións pragmáticas dificultan seriamente a súa adopción e utilidade dentro dos contornos organizacionais sociotécnicos. Máis aló da precisión, as solucións XAI deben demostrar que son utilizables, cumpridoras e eticamente aliñadas para garantir un impacto creíble.

No ámbito da minería de procesos, esta fenda maniféstase salientablemente a pesar do inmenso potencial das súas ideas baseadas en datos para transformar a eficiencia nos

procesos de negocio, así como nos procesos sociais críticos. As técnicas XAI actuais propostas para a minaría de procesos seguen estando estreitamente enfocadas, carecen de fondo cualitativo e están desconectadas dos ecosistemas reais de implantación.

En concreto, a revisión sistemática da literatura do Capítulo 2 revelou unha fixación predominante en métodos de atribución de características como SHAP para tarefas de monitorización predictiva, con menos estudos explorando técnicas de explicación causal ou contrastiva. Os beneficios hipotetizados das explicacións causais para dilucidar as dependencias entre eventos e explicacións contrastivas para clarificar a varianza dos arquetipos seguen sen aproveitarse en tarefas de descubrimento, análise e mellora de procesos. Esta sobrecarga na relevancia das características post hoc corre riscos de opacidade se os modelos complexos resisten a simplificación e carece de poder explicativo máis aló da puntuación de predicións.

As metodoloxías de avaliación predominantes tamén demostran deficiencias ao enfatizar métricas cuantitativas sobre avaliacións cualitativas de usuarios reais, exacerbación de puntos cegos con respecto á factualidade, comprensibilidade e relevancia. A maioría das técnicas XAI propostas carecen de base dentro de contextos de implantación organizacional a través de extensas probas de usuario que abranguen diversos roles, fluxos de traballo e sectores.

Apenas o 33% dos artigos analizados implementaron solucións PM explicables dentro de interfaces interactivas, ferramentas ou demostracións máis aló de exposicións teóricas. Unha investigación mínima examinou os retos organizacionais reais en torno á xestión do cambio, á privacidade, á seguridade, ós prexuízos e ós impactos en traballadores diversos que limitan a adopción do XAI. E apenas hai estudos que explorasen a integración da capacidade de explicación nos ciclos de vida do desenvolvemento de modelos, máis aló de complementos limitados desvinculados de procesos de adestramento cheos de riscos de opacidade

Esta rede enmarañada de deficiencias técnicas, pragmáticas e éticas nas solucións XAI existentes adaptadas para a minaría de procesos expón a inmadurez do campo a pesar das aspiracións infladas. O problema fundamental reside na desconexión entre a precisión prometida e a adopción lograda de explicacións que demostren ser realmente utilizables, cumpridoras e eticamente aliñadas. Xa que as regulamentacións de todo o mundo impoñen mandatos de transparencia algorítmica, esta deficiencia faise aguda dentro da pegada empresarial en expansión da minaría de procesos.

As leis emerxentes como a Lei de IA da UE requiren unha documentación exhaustiva de todo, desde as prácticas de datos ata os algoritmos para demostrar o cumprimento na implantación de sistemas de minaría de procesos de alto risco. Pero

as lagoas permiten eludir os estándares concretos de explicabilidade que benefician aos usuarios finais, especialmente dadas as lexítimas protección de segredo en torno a modelos propietarios. Así que as extensas cargas de supervisión pasan aos provedores sen garantir un valor significativo para aqueles que confían nos sistemas diariamente.

Incluso as regulacións evitan prescrever estándares técnicos, reflectindo a xenerosa flexibilidade da minaría de procesos que dificulta a gobernanza uniforme. Adapta-la supervisión ao amplo alcance metodolóxico que abrangue desde a descuberta de procesos ata a monitorización predictiva segue sendo ambigua sen orientación específica do contexto. E persisten as tensións para equilibra-los requisitos que garanten a rendición de contas externa a través da participación evitando ao mesmo tempo rigidices innecesarias que dificulten a innovación.

As implicacións éticas desta néboa regulatoria medran xunto coa permeación organizacional acelerada da minaría de procesos. As preocupacións sobre privacidade, responsabilidade, prexuízos e riscos de exclusión permanecen en gran medida sen reconciliar, faltando mecanismos de avaliación participativa que teñen en conta as perspectivas de usuarios e grupos heteroxéneos afectados por sistemas algorítmicos. Tanto as intervencións técnicas coma as probas de prexuízos, así como as iniciativas organizacionais e as xuntas de supervisión inclusivas, requiren unha integración que segue sendo escasa na práctica.

En resumo, sen dúbida lográronse avances considerables con respecto á precisión técnica nos algoritmos de minaría de procesos explicables ausente a validación centrada nas persoas dentro de contornas de implementación de destino. Pero este avance unilateral arrisca fracasar no limiar do impacto transformacional sen enfrontar as restricións do mundo real omnipresentes. A falta de profundidade cualitativa, o baleiro de avaliación, a complexa implementación organizacional opaca, a porosidade regulatoria e a fragilidade ética caracterizan o dilema central do campo.

Salvar este abismo require elevar a capacidade de explicar a minaría de procesos mediante enfoques interdisciplinarios completos que abranguen partes interesadas desde técnicos de taller ata liderado do consello. Impregnar innovacións técnicas cunha base ética, resonancia pragmática e coherencia de colaboración segue sendo imperativo aínda non realizado a través de disposicións regulamentarias, xestión da mudanza organizacional e prácticas do ciclo de vida máis alá da precisión conveniente. Un marco unificado que ataque as causas sistémicas do problema pode cataliza-lo salto cara a solucións sustentables centradas finalmente nos valores humanos a pesar da crecente complexidade tecnolóxica co tempo. Tendo conciencia deste momento crucial, un camiño adiante construtivo agora require reflexión.

## Contribucións

Esta investigación ten como obxectivo avanzar en técnicas de minaría de procesos explicables que demostren non só precisión senón tamén aliñamento integral con restricións do mundo real que abranguen usabilidade, cumprimento e ética. Contribúe con perspectivas integrais que pontean solucións conceptuais XAI coa realización pragmática dentro de ambientes organizacionais.

Este marco multidimensional transcende as contribucións técnicas illadas ao situar innovacións XAI dentro do contexto sociotécnico de procesos de negocio e restricións. Representa unha estrutura seminal que destaca necesidades, describe direccións e permita unha mellora continua á medida que os métodos e ecosistemas coevolucionan.

Ademais, catro estudos adicados proporcionan base empírica, xuntos abordando os ámbitos interrelacionados do problema investigado. Unha revisión sistemática da literatura analiza as técnicas XAI existentes propostas para a minaría de procesos, revelando tendencias en métodos causais e contrastivos que pasaron desapercibidos por unha abrumadora dependencia de explicacións baseadas en atribución de características para tarefas predictivas. Establece coñecementos básicos sobre a fenda de adopción-precisión.

A investigación cualitativa a través de enquisas e entrevistas descobre estratexias e barreiras do mundo real cos que se encontran os profesionais explicando ideas de proceso aos clientes. Os resultados revelan enfoques de entrega non estruturados xunto con obstáculos na integración de datos, suxerindo desaliñamentos organizacionais que dificultan a credibilidade. Xuntos, os coñecementos técnicos e prácticos fundidos expoñen a mocidade da capacidade de explicación na minaría de procesos a pesar das aspiracións infladas.

Análises separadas de regulamentos emerxentes e riscos éticos detallan a continuación as facetas adxacentes que conforman o progreso responsable XAI. As políticas da UE esixen unha ampla documentación para demostrar a responsabilidade algorítmica en sistemas de minaría de procesos de alto risco. Pero os dereitos individuais porosos de capacidade explicativa aínda carecen de estándares concretos que beneficien aos usuarios finais. Polo tanto, as cargas de supervisión pásanse aos provedores sen requirir un valor significativo para os destinatarios diarios do sistema.

Unha taxonomía de riscos que cataloga vulnerabilidades técnicas sobre prexuízos e privacidade xunto con preocupacións sociotécnicas sobre seguridade e argumentación establece unha base ética. Adaptar os riscos destacados aos contextos de minaría de procesos descobre tensións salientables e inexploradas sobre privacidade, responsabil-

idade e marxinación que requiren deliberación. Sintetizando análises regulamentarias e éticas remachamos o caso de solucións elevadas máis alá das métricas de precisión convenientes.

En conxunto, as investigacións técnicas, prácticas, legais e éticas interconectadas sustentan a arquitectura conceptual proposta para aliñar por fin as técnicas XAI innovadoras co impacto do despreguemento creíble. A combinación dunha base empírica e dunha orientación estrutural significa un progreso multifrontal cara ao esquivo obxectivo da minaría de procesos explicable que demostre non só precisión senón tamén integral á adopción.

Esta investigación tamén produce contribucións discretas como directrices de avaliación XAI para minaría de procesos e ferramentas que axudan á elaboración participativa de requirimentos de capacidade explicativa. A análise comparativa de políticas descubriu riscos éticos sobre responsabilidade opaca e impactos marxinalizantes que requiren compensación. A formulación de explicacións que coinciden cos modelos mentais, a xestión de expectativas e a facilitación da aliñación participativa foron descubertas como factores de éxito explicativo do mundo real.

En medio da incerteza regulatoria, xa que as xurisdicións navegan entre aspiracións de principios pragmáticos e estandarización pragmática, a dinámica competitiva tamén chama a atención. A reticencia da industria á transparencia derivada do segredo, a responsabilidade ou as preocupacións reputacionais xa enfangan a secuencia construtiva da xestión do cambio para as organizacións que adoptan sistemas de minaría de procesos. Pero unha transparencia medida enmarcada que enfatice as necesidades participativas sobre as exposicións punitivas pode asegurar a aceptación. Aínda así, require unha supervisión responsable enriquecendo as orientacións dos avaliadores desde a elaboración de listaxes de cumprimento ata o apoio consultivo holístico axustado ás intricacións modernas sobre a ética. A actualización de habilidades a través de compromisos da comunidade mundial demostrarase fundamental.

## Descubrimentos

As investigacións técnicas, prácticas, legais e éticas interconectadas realizadas producen varios descubrimentos clave que informan unha comprensión holística das limitacións nas técnicas actuais de minaría de procesos explicable:

A revisión sistemática da literatura en capítulo 2 analizou exhaustivamente o estado da arte das técnicas de XAI aplicadas á PM. A análise de 45 estudos relevantes revelou varias tendencias, contribucións influentes, limitacións e lagoas na investigación. En

canto ás tarefas de PM, a maioría dos estudos centrouse no seguimento preditivo de procesos, especialmente na predición de resultados e eventos seguintes. As técnicas XAI máis empregadas foron os métodos de atribución de características, como SHAP e LIME, para explicar as decisións dos modelos. Non obstante, as técnicas de explicación causal permanecen relativamente infrautilizadas máis alá das tarefas de predición. Ademais, a maioría dos estudos dependen en gran medida de conxuntos de datos de referencia públicos en lugar de rexistros de organizacións do mundo real. Isto limita a avaliación da utilidade práctica das técnicas XAI en entornos aplicados. A revisión tamén puxo de manifesto a falta dunha avaliación cualitativa das interfaces de explicación e das barreiras organizativas que poden impedir a adopción exitosa das innovacións XAI. A literatura céntrase abrumadoramente na precisión técnica en lugar da utilidade pragmática aliñada coas diversas necesidades dos usuarios e os contextos de implantación.

El estudo cualitativo en capítulo 3 parte dos resultados da revisión sistemática analizada no capítulo anterior. Mentres que a SLR se centrou en analizar o estado da arte nas técnicas de XAI aplicadas á PM, este estudo investiga as estratexias e barreiras reais para explicar as perspectivas da PM dende a perspectiva dos profesionais. O estudo empregou unha metodoloxía en dúas fases, que consistiu nun cuestionario en liña e entrevistas a grupos focais. O cuestionario tiña como obxectivo recoller datos sobre as estratexias empíricas actualmente utilizadas polos profesionais, mentres que as entrevistas permitiron aos participantes ampliar os seus puntos de vista e proporcionar perspectivas máis profundas.

Os resultados clave do estudo suxiren que a explicabilidade da inspección de procesos percíbese como necesaria, pero as estratexias de entrega non están claras. Os profesionais destacaron a importancia de proporcionar explicacións accesibles e obxectivas ás partes interesadas non técnicas, adoptando un enfoque diplomático e fomentando a colaboración e a reflexión. As barreiras á integración de datos e sistemas dificultan a interpretabilidade, sendo crucial unha integración eficaz de datos e sistemas para ofrecer explicacións precisas e informadas aos clientes. Os profesionais recomendaron priorizar as actividades relevantes, involucrar ao persoal de TI e enxeñeiros de datos desde o inicio do proceso e establecer robustas instalacións de preprocesamento. Ademais, as estratexias de explicabilidade dependen do contexto, o que fai que a xeneralización sexa un reto. A eficacia das estratexias de explicabilidade está influenciada polo contexto único da organización e do proxecto de cada cliente. Os profesionais suxeriron avaliar a cultura organizativa, proporcionar historias de referencia, diferenciar entre a PM como obxectivo e como método, e adaptar as

explicacións ás intencións dos usuarios e aos seus coñecementos técnicos. Por outra banda, existe unha fenda entre a madurez dos clientes e a demanda de explicacións modulares. As estratexias para abordar esta fenda inclúen proporcionar formación e sesións introdutorias sobre PM, involucrar aos clientes nos percorridos e validacións de procesos, filtrar a información para maximizar a limitada atención dos clientes e realizar avaliacións exhaustivas da madurez da xestión de procesos empresariais.

Os resultados deste estudo de profesionais proporcionan perspectivas valiosas sobre como se abordan os retos identificados na SLR en contornas reais. O estudo destaca a importancia de desenvolver técnicas XAI que sexan precisas, escalables e adaptables ás diferentes necesidades dos usuarios e aos contextos organizativos. Tamén subliña a necesidade de marcos de avaliación centrados no usuario, métodos XAI específicos de cada dominio e enfoques con múltiples partes interesadas para apoiar a adopción de iniciativas explicables de PM.

Os capítulos 4 e 5 amplían os resultados da SLR e do estudo cualitativo, examinando as posturas reguladoras e éticas sobre a explicabilidade na minería de procesos, con especial atención ao contexto da UE. O Capítulo 4 analiza o panorama regulador internacional relacionado coa explicabilidade da IA, revelando desafíos comúns e diferenzas estratéxicas entre rexións. Céntrase especialmente nos regulamentos da UE que esixen formas de explicabilidade para os sistemas algorítmicos, catalogando as disposicións legais nunha taxonomía estruturada. Esta análise detallada xustifícase polo papel destacado da UE na gobernanza da IA e pola concentración de actividades de minería de procesos en Europa, como evidencian os resultados do cuestionario. O estudo destaca as tensións entre os desexos de transparencia e as barreiras á divulgación completa, ilustrando os problemas de forma concreta para a minería de procesos a través dun escenario de exemplo.

O Capítulo 5 aborda as consideracións éticas e os riscos que deben terse en conta ao aplicar técnicas de XAI no contexto da minería de procesos. Sintetiza ideas de teorías éticas, marcos de ética aplicada e estudos de casos da industria para sacar á luz os desafíos e tensións específicos do contexto que poden xurdir cando se procura a transparencia dos procesos. O capítulo ofrece unha taxonomía de riscos técnicos e sociotécnicos asociados coa implantación de técnicas de XAI na minería de procesos, abarcando dimensións como a robustez, a xustiza, a privacidade e os factores sociotécnicos. Tamén presenta un conxunto de principios éticos e estratexias de mitigación de riscos que poden orientar o desenvolvemento e a implantación responsables de XAI na minería de procesos.

En conxunto, estes capítulos amplían a análise previa, subliñando a importancia

de considerar as implicacións reguladoras e éticas da explicabilidade na minería de procesos. Resaltan a necesidade de desenvolver técnicas de XAI que non só sexan precisas e escalables, senón tamén adaptables ás diferentes necesidades dos usuarios e aos contextos organizativos. Ademais, recalcan a importancia de establecer marcos de avaliación centrados no usuario, métodos XAI específicos de cada dominio e enfoques con múltiples partes interesadas para apoiar a adopción de iniciativas explicables de minería de procesos.

O Capítulo 6 presenta un marco conceptual integral para a explicabilidade na minería de procesos, que aborda os desafíos e as limitacións identificadas nos capítulos anteriores. O marco propón un enfoque por fases que abarca a obtención de requisitos, o desenvolvemento de ferramentas técnicas, a integración organizativa e as estruturas de gobernanza e políticas. O marco comeza definindo dimensións clave de explicabilidade, como obxectivos, tipos, contextos e persoas, para orientar o espazo de solucións. A seguir, detalla un enfoque por fases que inclúe a obtención de requisitos impulsada polas partes interesadas, o desenvolvemento responsable de ferramentas técnicas, os procesos de integración organizativa e as estruturas de gobernanza e políticas internas e externas.

O Capítulo 7 ilustra a aplicación do marco de explicabilidade nun estudo de caso hipotético no ámbito sanitario. O estudo céntrase na optimización do proceso para pacientes con enfermidades valvulares cardíacas nun entorno hospitalario. Demostra como as técnicas de minería de procesos poden utilizarse para analizar e mellorar o complexo procedemento cirúrxico, dende o diagnóstico inicial ata o coidado postoperatorio. O capítulo describe a implementación do marco de explicabilidade en seis pasos, abordando desafíos como a minimización do tempo de decisión, a xestión de complicacións dos pacientes, o cumprimento dos acordos de nivel de servizo e a extracción de información accionable a partir dos datos do proceso. O marco adáptase ao contexto sanitario, considerando as necesidades específicas dos clínicos, pacientes e administradores hospitalarios, á vez que equilibra os requisitos de explicabilidade cos posibles riscos e custos.

O Capítulo 8 conclúe a tese sintetizando os achados dos capítulos anteriores para abordar a pregunta principal de investigación sobre a integración de técnicas XAI e procesos de deseño participativo nos sistemas de PM. O capítulo tamén esboza prometedoras direccións de investigación futura, como o desenvolvemento de explicacións multidimensionais e sensibles ao contexto, o avance da inferencia causal e as explicacións contrafactuais en PM, a creación de interfaces de linguaxe natural para os insights de PM, a exploración de técnicas de explicación que preserven a privacidade,

o aproveitamento da simulación para a validación de explicacións e o establecemento de marcos de gobernanza ética da IA específicos para PM.

En resumo, esta tese aborda os desafíos multifacéticos da explicabilidade na minería de procesos desde perspectivas técnicas, empíricas, regulatorias e éticas. En xeral, a tese contribúe ao avance do campo da minería de procesos explicable proporcionando unha análise interdisciplinaria dos desafíos, un marco práctico para a implementación responsable e unha axenda para a investigación e mellora futura.

# Summary

In an era where data-driven insights are shaping the future of organizations, Process Mining (PM) has emerged as a transformative force, offering unprecedented opportunities to analyze and optimize complex processes. However, the full potential of PM remains untapped due to persistent challenges in understanding and explainability of implemented process technology, such as AI systems. This thesis establishes a structured investigation to explore the role of Explainable AI (XAI) in overcoming these barriers, fostering greater adoption and engagement with PM solutions, and ultimately bridging the gap between advanced technologies and human understanding.

## Problem Statement

The rapid proliferation of artificial intelligence (AI) technologies like machine learning into critical domains has raised concerns about their transparency and accountability. In particular, complex AI systems employed in consequential operational processes pose threats stemming from their opacity, heightening risks of errors, biases, and harms without recourse. This underscores the crucial need for XAI techniques that render these AI models transparent and interpretable.

However, significant deficiencies persist regarding the rigorous application of XAI solutions in real-world practice. While technical XAI algorithms alone may claim to provide transparency, pragmatic constraints severely hinder their adoption and utility within sociotechnical organizational settings. Beyond accuracy, XAI solutions must demonstrate being usable, compliant, and ethically aligned to ensure credible impact.

In the realm of process mining, this gap manifests prominently despite the immense potential of its data-driven insights to transform efficiency in business processes as well as critical social processes. Existing XAI techniques proposed for process mining remain narrowly focused, lack qualitative grounding, and are disconnected from real deployment ecosystems.

Specifically, the systematic literature review in Chapter 2 revealed a predominant fixation on feature attribution methods like SHAP for predictive monitoring tasks, with fewer studies exploring causal or contrastive explanation techniques. The hypothesized benefits of causal explanations for elucidating dependencies between events and contrastive explanations for clarifying variance from archetypes remain untapped in process discovery, analysis, and improvement tasks. This overload on post-hoc feature relevance risks opacity if complex models resist simplification and lacks explanatory power beyond prediction scoring.

Prevailing evaluation methodologies also demonstrate shortcomings by emphasizing quantitative metrics over qualitative assessments with real users, exacerbating blind spots regarding factuality, comprehensibility, and relevance. Most proposed XAI techniques lack grounding within organizational deployment contexts through extensive user testing spanning diverse roles, workflows, and sectors.

Merely 33% of analyzed papers implemented explainable PM solutions within interactive interfaces, tools, or demonstrations beyond theoretical expositions. Minimal research has examined real-world organizational challenges around change management, privacy, security, biases, and impacts on diverse workers that constrain XAI adoption. And scarcely any studies have explored integrating explainability into model development lifecycles beyond narrow plug-ins divorced from opacity-fraught training pipelines.

This tangled web of technical, pragmatic, and ethical deficiencies in existing XAI solutions tailored for process mining exposes the field's immaturity despite inflated aspirations. The core problem lies in the disconnect between promised accuracy and achieved adoption of explanations proven to be truly usable, compliant, and ethically aligned. As regulations worldwide impose algorithmic transparency mandates, this shortcoming grows acute within process mining's expanding enterprise footprint.

Emerging laws like the EU AI Act require extensive documentation of everything from data practices to algorithms to demonstrate compliance in deploying high-risk process mining systems. But loopholes allow circumventing concrete explainability standards benefiting end-users, especially given legitimate trade secrecy protections around proprietary models. So extensive oversight burdens fall on providers without ensuring meaningful value for those relying on the systems daily.

Even regulations avoid prescribing technical standards, reflecting process mining's generous flexibility that hinders uniform governance. Tailoring oversight to the broad methodological scope spanning process discovery to predictive monitoring remains ambiguous without context-specific guidance. And tensions persist in balancing

requirements ensuring external accountability through engagement while avoiding unnecessary rigidities stifling innovation.

Ethical implications of this regulatory haze grow alongside process mining's accelerated organizational permeation. Concerns over privacy, accountability, biases, and exclusion risks remain largely unreconciled, lacking participatory assessment mechanisms that account for perspectives of heterogeneous users and groups impacted by algorithmic systems. Both technical interventions like bias testing and organizational initiatives like inclusive oversight boards require an integration that remains scarce in practice.

In summary, considerable strides have undoubtedly been made regarding technical accuracy in explainable process mining algorithms absent people-centric validation within target deployment environments. But this one-sided advance risks faltering at the threshold of transformative impact without confronting ubiquitous real-world constraints. Lack of qualitative depth, evaluation void, opaque organizational implementation intricacy, regulatory porousness, and ethical fragility characterize the field's central dilemma. Bridging this chasm requires elevating process mining explainability through comprehensive interdisciplinary approaches spanning stakeholders from workshop technicians to board leadership. Infusing technical innovations with an ethical foundation, pragmatic resonance, and collaborative coherence remains an unrealized imperative across regulatory provisions, organizational change management, and lifecycle practices beyond convenient accuracy. A unified framework tackling the problem's systemic roots can catalyze the leap towards sustainable solutions ultimately centered on human values despite increasing technological complexity over time.

## Contributions

This research aims to advance explainable process mining techniques that demonstrate not only accuracy but also holistic alignment with real-world constraints spanning usability, compliance, and ethics. It contributes comprehensive perspectives bridging conceptual XAI solutions with pragmatic realization within organizational settings.

This multidimensional framework transcends isolated technical contributions by situating XAI innovations within the sociotechnical context of business processes and constraints. It represents a seminal structure highlighting needs, outlining directions, and enabling continuous improvement as methods and ecosystems co-evolve.

Additionally, four dedicated studies provide empirical grounding, together addressing the interrelated facets of the investigated problem. A systematic literature

review analyzes existing XAI techniques proposed for process mining, revealing trends in causal and contrastive methods overlooked by an overwhelming reliance on feature attribution-based explanations for predictive tasks. It establishes foundational insights into the adoption-accuracy gap.

Qualitative research through surveys and interviews uncovers real-world strategies and barriers encountered by practitioners in explaining process insights to clients. Findings reveal unstructured delivery approaches alongside obstacles in data integration, suggesting organizational misalignments hampering credibility. Together, the merged technical and practical insights expose the infancy of explainability in process mining despite inflated aspirations.

Separate analyses of emerging regulations and ethical risks then detail the adjacent facets shaping responsible XAI progress. EU policies demand extensive documentation to demonstrate algorithmic accountability in high-risk process mining systems. But porous individual explainability rights still lack concrete standards benefiting end-users. Thus, oversight burdens shift to providers without mandating meaningful value for the system's daily recipients.

A risk taxonomy cataloguing technical vulnerabilities around bias and privacy alongside sociotechnical concerns about safety and fairness establishes an ethical foundation. Adapting highlighted risks to process mining contexts uncovers salient, unexplored tensions around privacy and accountability. Synthesizing regulatory and ethical analyses clinches the case for elevated solutions beyond convenient accuracy metrics.

Collectively, the interconnected technical, practical, legal, and ethical investigations undergird the proposed conceptual architecture to finally align innovative XAI techniques with credible deployment impact. The combination of empirical grounding and structural guidance signifies multifrontal progress towards the elusive goal of explainable process mining demonstrating not only accuracy but also integral to adoption.

This research also yields discrete contributions like XAI evaluation guidelines for process mining and tools assisting participatory explainability requirements elicitation. Policy benchmarking uncovered ethical risks around opaque accountability and marginalizing impacts necessitating mitigation. Formulating explanations matching mental models, managing expectations, and facilitating participatory alignment emerged as real-world explanatory success factors.

Amidst regulatory uncertainty as jurisdictions navigate between pragmatic aspirations of principled pragmatism and standardization, competitive dynamics also beckon

attention. Industry reticence towards transparency stemming from trade secrecy, liability, or reputational concerns already muddies constructive change management sequencing for organizations adopting process mining systems. But measured framed transparency emphasizing participatory needs over punitive exposures can ensure acceptance. Still, it requires responsible oversight enriching auditor guidance from compliance checklists to holistic consultative support attuned to modern intricacies around ethics. Skill upgrading through global community engagements will prove pivotal.

## Findings

The interconnected technical, practical, legal, and ethical investigations conducted yield several key findings informing a holistic understanding of limitations in current explainable process mining techniques:

The systematic literature review in Chapter 2 comprehensively analyzed the state-of-the-art XAI techniques applied to PM. Analysis of 45 relevant studies revealed several trends, influential contributions, limitations, and research gaps. Regarding PM tasks, most studies focused on predictive process monitoring, especially outcome and next event prediction. The most employed XAI techniques were feature attribution methods like SHAP and LIME to explain model decisions, while causal explanation techniques remain relatively underutilized beyond prediction tasks. Moreover, most studies heavily rely on public benchmark datasets rather than real-world organization logs. This limits the evaluation of XAI techniques' practical utility in applied settings. The review also highlighted the lack of qualitative evaluation of explanation interfaces and organizational barriers that can hinder successful XAI innovations adoption. Literature overwhelmingly focuses on technical accuracy over pragmatic utility aligned with diverse user needs and deployment contexts.

The qualitative study in Chapter 3 builds on the findings from the systematic review analyzed in the previous chapter. While the SLR focused on analyzing the state-of-the-art in XAI techniques applied to PM, this study investigates the real-world strategies and barriers to explaining PM insights from the perspective of practitioners. The study employed a two-phase methodology, consisting of an online questionnaire and focus group interviews. The questionnaire aimed to collect data on the empirical strategies currently used by practitioners, while the interviews allowed participants to expand on their views and provide deeper insights. Key findings from the study suggest that process mining explainability is perceived as necessary, but delivery strategies

are unclear. Practitioners emphasized the importance of providing accessible, objective explanations to non-technical stakeholders, adopting a diplomatic approach, and fostering collaboration and reflection. Barriers to data and system integration hinder interpretability, with effective data and system integration being crucial for delivering accurate, informed explanations to clients. Practitioners recommended prioritizing relevant activities, engaging IT staff and data engineers early in the process, and establishing robust preprocessing pipelines. Moreover, explainability strategies are context-dependent, making generalization challenging. The effectiveness of explainability strategies is influenced by each client's unique organizational and project context. Practitioners suggested assessing organizational culture, providing reference stories, differentiating between PM as a goal and as a method, and tailoring explanations to user intentions and technical proficiency. Furthermore, there is a gap between client maturity and the demand for modular explanations. Strategies to address this gap include providing PM training and introductory sessions, engaging clients in process walkthroughs and validations, filtering information to maximize clients' limited attention spans, and conducting thorough business process management maturity assessments. The findings from this practitioner study provide valuable insights into how the challenges identified in the SLR are addressed in real-world settings. The study then highlights the importance of developing XAI techniques that are accurate, scalable, and adaptable to different user needs and organizational contexts. It also underscores the need for user-centric evaluation frameworks, domain-specific XAI methods, and multi-stakeholder approaches to support the adoption of explainable PM initiatives.

Chapters 4 and 5 extend the findings from the SLR and qualitative study by examining regulatory and ethical stances on explainability in process mining, with a particular focus on the EU context. Chapter 4 analyzes the international regulatory landscape related to AI explainability, revealing common challenges and strategic differences across regions. It especially focuses on EU regulations mandating forms of explainability for algorithmic systems, cataloguing legal provisions into a structured taxonomy. This detailed analysis is justified by the EU's prominent role in AI governance and the concentration of process mining activities in Europe, as evidenced by questionnaire findings. The study highlights tensions between transparency desires and barriers to full disclosure, illustrating issues concretely for process mining through an example scenario.

Chapter 5 addresses the ethical considerations and risks that need to be accounted for when applying XAI techniques in the context of process mining. It synthesizes

insights from ethical theories, applied ethics frameworks, and industry case studies to surface context-specific challenges and tensions that may arise when pursuing process transparency. The chapter offers a taxonomy of technical and socio-technical risks associated with deploying XAI techniques in process mining, spanning dimensions such as robustness, fairness, privacy, and socio-technical factors. It also presents a set of ethical principles and risk mitigation strategies that can guide the responsible development and deployment of XAI in process mining.

Collectively, these chapters extend the prior analysis, underscoring the importance of considering the regulatory and ethical implications of explainability in process mining. They highlight the need to develop XAI techniques that are not only accurate and scalable but also adaptable to different user needs and organizational contexts. Additionally, they emphasize the importance of establishing user-centric evaluation frameworks, domain-specific XAI methods, and multi-stakeholder approaches to support the adoption of explainable process mining initiatives.

Chapter 6 presents a comprehensive conceptual framework for explainability in process mining, addressing the challenges and limitations identified in the previous chapters. The framework proposes a phased approach spanning requirements elicitation, technical tool development, organizational integration, and governance and policy structures. The framework begins by defining key dimensions of explainability, such as objectives, types, contexts, and personas, to guide the solution space. It then details a phased approach including stakeholder-driven requirements elicitation, responsible technical tool development, organizational integration processes, and internal and external governance and policy structures.

Chapter 7 illustrates the application of the explainability framework in a hypothetical healthcare case study. The study focuses on optimizing the process for patients with heart valve diseases in a hospital setting. It demonstrates how process mining techniques can be used to analyze and improve the complex surgical procedure, from initial diagnosis to post-operative care. The chapter outlines the implementation of the six-step explainability framework, addressing challenges such as minimizing decision time, managing patient complications, ensuring compliance with service level agreements, and extracting actionable insights from process data. The framework is adapted to the healthcare context, considering the specific needs of clinicians, patients, and hospital administrators while balancing explainability requirements with potential risks and costs.

Chapter 8 concludes the thesis by synthesizing the findings from previous chapters to address the main research question on integrating XAI techniques and participa-

tory design processes into PM systems. The chapter also outlines promising future research directions, such as developing multi-perspective and context-aware explanations, advancing causal inference and counterfactual explanations in PM, creating natural language interfaces for PM insights, exploring privacy-preserving explanation techniques, leveraging simulation for explanation validation, and establishing ethical AI governance frameworks specific to PM.

In summary, this thesis addresses the multifaceted challenges of explainability in process mining from technical, empirical, regulatory, and ethical perspectives. Overall, the thesis contributes to advancing the field of explainable process mining by providing an interdisciplinary analysis of challenges, a practical framework for responsible implementation, and an agenda for future research and improvement.

# Contents

# CHAPTER 1

# INTRODUCTION

## 1.1 Motivation

In recent years, Process Mining (PM) has rapidly gained traction as a way to increase transparency and optimize organizational processes across diverse industries including healthcare, manufacturing, and IT services [399,467,469]. By extracting valuable insights from event log data, PM enables evidence-driven process enhancement. However, amid the complexity of the techniques involved, PM solutions often suffer from a lack of explainability that hinders stakeholder comprehension and engagement.

This issue is further compounded by the integration of opaque artificial intelligence systems into process analytics, which can exacerbate distrust in model outputs. Techniques like deep learning, despite their predictive power, act as "black boxes" that do not reveal the underlying logic behind their functioning. The incorporation of AI hence makes the inner workings of PM even more obscure to non-technical experts within organizations. Consequently, there is a need to explore how AI explainability can help demystify PM, tailoring it to human understanding and domain knowledge. However, current literature in both academia and industry provides limited guidance on explainable PM that considers the organizational context. Research on explainable AI (XAI) itself is an evolving field, with open debates around key questions of what makes an explanation useful and how to evaluate explanation quality [35,180,269,325,326].

Therefore, this thesis aims to conduct an extensive analysis into XAI techniques and their applicability in making process analytics transparent, interpretable, and actionable. The scope encompasses both technical dimensions, like algorithms and data, alongside human factors, organizational culture, and real-world constraints. The intended contribution is a comprehensive XAI framework tailored to PM across the

dimensions of usability, auditability, and ethical alignment. The framework integrates multidisciplinary perspectives spanning artificial intelligence, visual analytics, communication theory, management science, and design. By elucidating the connections between advanced process analytics and human values, this research can enable organizations to tap into the full potential of PM in a responsible manner. The findings can guide technology vendors in developing transparent solutions while empowering practitioners with concrete strategies for explainable process enhancement.

PM is predicated on a set of foundational concepts that allow for the systematic collection, analysis, and representation of process data. These building blocks—events, traces, logs, and process models—serve to contextualize the myriad activities within a process. As we delineate these elements, we will also elaborate on different notational methods used to represent processes, and the core activities that constitute the PM domain.

**Definition 1: Event**

An event $\epsilon$ serves as an atomic unit in PM, capturing the occurrence of a singular activity within a specific process instance. Formally, an event $\epsilon$ is constituted by a 4-tuple $(\alpha, t, c, \phi)$:

- $\alpha \in A$: Specifies the executed activity, with $A$ encompassing the universe of activities germane to the process under scrutiny.
- $t \in \mathbb{T}$: The timestamp $t$ demarcates the temporal instance of the event's occurrence within a well-ordered set $\mathbb{T}$.
- $c \in \mathbb{C}$: Represents a case identifier, uniquely ascribing the event to a specific process instance.
- $\phi : \mathbb{K} \rightarrow \mathbb{V}$: An attribute function that encapsulates auxiliary contextual attributes by mapping keys $\mathbb{K}$ to values $\mathbb{V}$.

Events are the most granular level of data in PM, encapsulating a single occurrence of an activity within a given process. They serve as the cornerstone for more complex structures like traces and logs. The formal definition of an event incorporates not just the activity, but also temporal data, instance identification, and contextual attributes.

**Definition 2: Trace**

A trace $\tau$ is an ordered sequence of events aggregated under the same process instance.

It is formally represented as a function $\tau : \mathbb{N} \to E$, such that:

$$\forall i, j \in \mathbb{N}, (i < j) \Rightarrow (\tau(i).c = \tau(j).c) \wedge (\tau(i).t < \tau(j).t)$$

Here, $E$ denotes the totality of events.

Traces elevate the granularity of process analysis by aggregating individual events under the umbrella of a single process instance. In doing so, they capture the sequential dependencies and transitional probabilities between activities. Traces are more than mere collections of events; they represent the chronological unfolding of events within a specific context, thereby facilitating the study of process pathways and variations.

**Definition 3: Event Log**

An event log $L$ acts as a multiset containing a collection of traces $\tau$, subject to the uniqueness constraint:

$$\forall \tau(i), \tau(j) \in L, i \neq j \Rightarrow \tau(i).c \neq \tau(j).c$$

Event logs are the aggregate records that serve as the primary input for PM algorithms. Comprising multiple traces, each corresponding to a unique process instance, logs provide a macroscopic view of the process landscape. They are instrumental for deriving process models through discovery techniques, validating them through conformance checking, and refining them via enhancement methods.

**Definition 4: Process Model Notational Representations**

Process Model Notational Representations serve as formal frameworks for depicting the structure and behavior of processes. Based on [467], common notations in PM are:

1. **Petri Net:** A Petri Net $N$ is formalized as a 4-tuple $(P, T, F, M)$, where $P$ and $T$ are finite, disjoint sets of places and transitions, respectively; $F$ comprises directed arcs; and $M : P \to \mathbb{N}$ serves as the marking function. Petri Nets are useful for capturing process concurrency and synchronization aspects.
2. **Directly-Follows Graph (DFG):** A Directly-Follows Graph (DFG) is a directed graph $G = (V, E)$, where $V$ signifies activities and $E$ denotes the direct-follows relation between them. DFGs are effective for capturing the sequential dependencies between activities in a simplified manner.
3. **Business Process Model and Notation (BPMN):** BPMN, or Business Process Model and Notation, is a standardized graphical representation that employs

a comprehensive set of symbols and notations for depicting complex business processes. BPMN models can capture a wide range of process aspects, including data flows, resource allocation, and exception handling.

4. **Process Trees:** A Process Tree is a hierarchical representation of a process model, where leaf nodes represent activities and internal nodes represent control-flow operators (e.g., sequence, choice, parallel). Process Trees are advantageous for their sound-by-construction property and their ability to represent block-structured processes.

The representation of processes is a critical task that demands a nuanced approach capable of capturing the multifaceted nature of business processes. Different notational schemes like Directly-Follows Graphs (exemplified in Figure 1.1), Petri Nets, BPMN (Figure 1.2), and Process Trees cater to various needs—be it the detailed representation of control flows, the simplification of complex systems, or the visualization of business processes. These notations are not mutually exclusive and are often used in tandem for comprehensive process analysis.

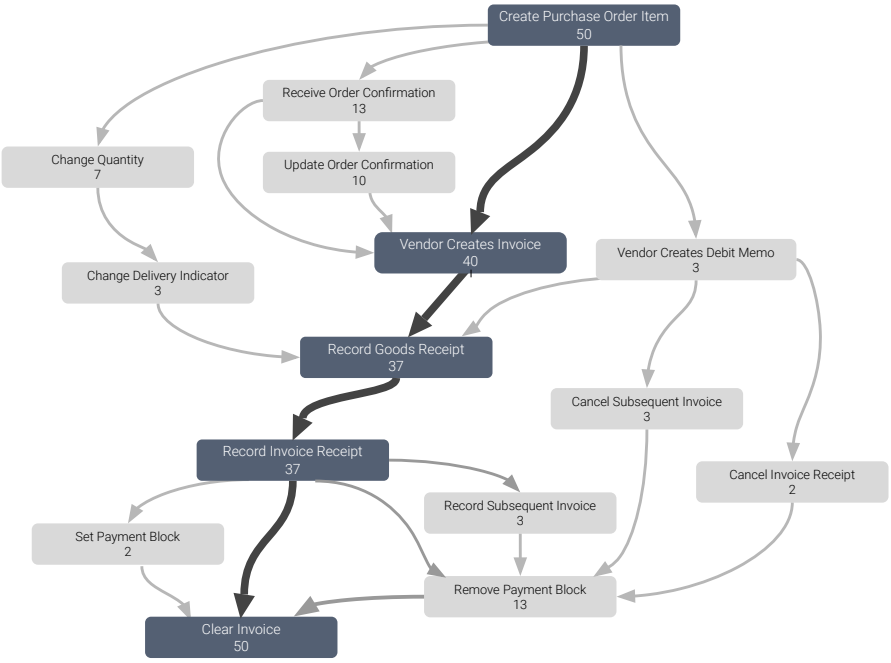


**Figure 1.1:** A Directly-Follows Graphs (DFGs) model of a purchase order (PO) process, inspired from the BPI challenge 2019 dataset (474).

**Definition 5: Core Activities in PM**

Core activities in PM encompass a range of techniques that are pivotal for extracting valuable insights from event logs. These can be broadly categorized into the following:

- *Process discovery* - Automatically constructing process models from event logs to represent the observed behavior. Process discovery algorithms, such as the Alpha algorithm [470], region-based approaches [71, 441, 470, 473], inductive mining [272, 273], and the split miner [37], can generate models in various notations, including Petri nets, BPMN, and process trees. These models capture the process flow, including sequential, choice, parallel, and loop constructs, providing valuable insights into the actual process execution [468].
- *Conformance checking* - Comparing the observed behavior in the event log against a reference process model to identify commonalities and discrepancies. Conformance checking techniques [72] can detect deviations from normative models or exceptional cases that deviate from automatically discovered models, helping identify process inefficiencies.
- *Predictive process monitoring* - Leveraging event data and process models to predict future process outcomes, such as the remaining processing time, the likelihood of deadline violations, or the probability of a case following a specific path [157, 299]. Predictive process monitoring extends performance analysis – i.e. such as strengthening process model analyzing activity durations, waiting times, and resource utilization, to diagnose bottlenecks and inefficiencies – by using historical data to make predictions about ongoing or future cases.
- *Process enhancement and diagnostics* - Improving processes based on the insights gained from PM and identifying the root causes of performance and compliance issues [99]. Process enhancement techniques, such as simulation, optimization, and redesign, can be applied to suggest process improvements. Diagnostic techniques, such as root cause analysis and process deviation analysis, help understand the factors contributing to process inefficiencies and non-compliance.

Interpretability and explainability are crucial aspects of PM, as the insights and recommendations generated must be understandable and actionable for process stakeholders. In PM, interpretability often relates to the comprehensibility of the discovered process models [65]. Notations like Petri nets, BPMN, and process trees provide a range of constructs to represent process behavior, allowing for the creation of models that balance accuracy and understandability [468]. Explainability, on the other hand,

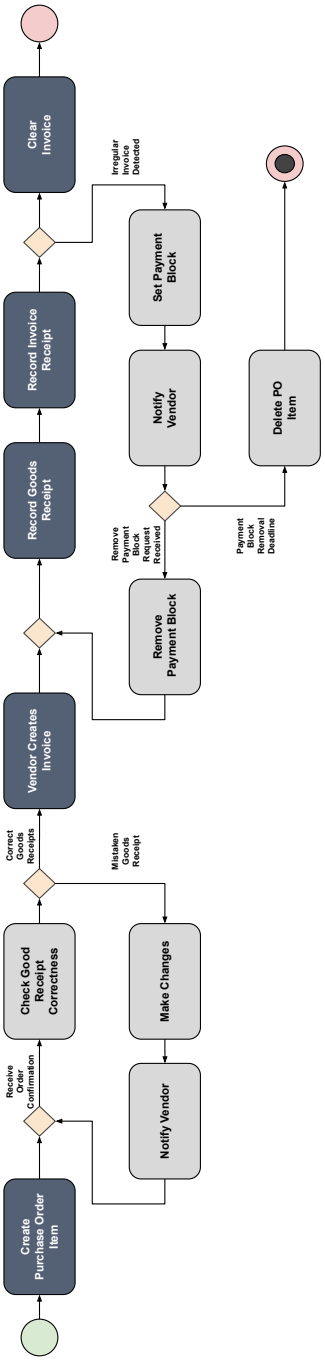


**Figure 1.2:** The same PO process represented in Figure 1.1 but via BPMN.

relates to justify the insights and recommendations derived from PM techniques, e.g., conformance checking results to clearly highlight and explain the reasons behind detected deviations or process enhancement recommendations accompanied by transparent justifications based on the analyzed event data and process models.

Nevertheless, as PM techniques become more sophisticated, ensuring interpretability and explainability becomes more challenging. Advanced process discovery algorithms, such as those based on region theory or inductive mining, may generate complex models that are difficult for stakeholders to comprehend [412]. Foremost, the application of machine learning techniques to PM tasks can lead to opaque models and predictions [304, 313].

### 1.1.1 Explainable AI Techniques

As AI models become increasingly sophisticated and opaque, the need for techniques that can shed light on their inner workings and elucidate their decision-making processes has become paramount. Explainable AI (XAI) seeks to develop methods and frameworks for making AI systems more transparent, interpretable, and accountable [180,366]. Despite a plethora of surveys and taxonomies now available [293,414], the field of XAI encompasses a wide range of techniques [5, 35, 178] that can be categorized as follows:

**Definition 6: Approaches in Explainable AI**

- *Intrinsic transparent models* are inherently interpretable due to their simple and transparent structures. These models include rule learners such as decision lists [322] and Ripper [84], linear models like regression [193] and support vector machines (SVM) [92], tree models such as decision trees and model trees [250, 392], and Bayesian models like Naive Bayes and Bayesian networks [280,385]. These models allow users to directly understand the decision-making process without requiring additional explanations [242, 413].
- *Model-specific post-hoc* methods provide explanations by examining the internal workings of a specific AI model. Attention mechanisms [40, 479] have been widely used to highlight the importance of input features in deep learning models. Layer-wise relevance propagation [331] attributes the output of a neural network to its input features by backpropagating the relevance scores. Structural equation models [386] and concept activation vectors [243] are other techniques that offer insights into the relationships between variables and the activation of high-level concepts in deep learning models.

- *Model-agnostic post-hoc* methods can be applied to any AI model, regardless of its architecture. Perturbation-based techniques, such as LIME (Local Interpretable Model-Agnostic Explanations) [402], Shapley values [296], and Anchors [403], generate explanations by perturbing the input data and observing the model's output. Gradient-based techniques, including Integrated Gradients and SmoothGrad [438, 453], analyze the model's gradients to identify the importance of input features. Counterfactual explanations [447] provide insights by generating examples that are minimally different from the original input but result in a different model prediction. Prototype selection methods [242, 281] identify representative instances that are most similar to the input data. Surrogate models (e.g. LIME) approximate the behavior of a complex model with a simpler, interpretable model.
- *Visual explanations* offer graphical representations of the relationship between input features and model predictions. Partial dependence plots [31, 159] show the marginal effect of a feature on the model's output, while individual conditional expectations [173] visualize the model's behavior for individual instances. Saliency maps [425, 431] highlight the importance of input features by computing the gradient of the model's output with respect to the input. Activation maximization [124] generates input patterns that maximize the activation of a specific neuron or layer in a deep learning model. while dimensionality reduction techniques (e.g., t-SNE and PCA [224, 471]), can be used to visualize high-dimensional data in a lower-dimensional space, revealing patterns and clusters.
- *Local explanations* focus on providing explanations for individual instances (instance-level) [295, 403], subgroups of instances (subgroup-level) [13, 79], or specific segments of data (log slice-level) [475]. These explanations offer a more granular understanding of the model's behavior for particular cases or scenarios, enabling users to gain insights into the factors influencing the model's decisions.

## 1.2 Hypotheses and Objectives

Fundamentally, this research aims to tackle a pivotal problematic in leveraging PM's immense potential - the inability of current XAI techniques to foster intuitive comprehension that resonates with the needs of diverse stakeholders spanning operational experts to executive leadership. While technical capabilities have achieved predictive gains on benchmarks, pragmatic barriers around usability and ethical alignment sever

realization within complex organizational ecosystems [246, 313, 332, 508]. The core question we pose is:

> ***How can explainable AI (XAI) techniques and participatory design processes be effectively integrated into PM systems to improve understanding, foster adoption, and ensure solutions that are accurate, usable, compliant and ethically aligned?***

The emphasis lies in elevating XAI from fragmented proofs-of-concept to responsible transformations attuned to usability constraints and ethical values. Four interconnected hypotheses motivate a constructive path forward:

**H1.** Current XAI techniques in PM are technically proficient but lack pragmatic utility and ethical alignment needed for organizational adoption.

**H2.** Explanations must resonate with mental models of diverse users, not just showcase accuracy on benchmarks.

**H3.** Participatory design can elicit needs and constraints to shape human-centered XAI solutions.

**H4.** Responsible innovation requires going beyond technical achievements to address usability, compliance, transparency.

The primary objective of this research is to establish a holistic, multidimensional framework for XAI in PM —a framework that addresses the complexities of regulatory landscapes, acknowledges the diversity of stakeholder needs, and provides robust mechanisms for the evaluation of explainability efficacy. This objective is underpinned by an interdisciplinary approach, drawing insights from AI ethics, human-computer interaction, and the legal dimensions of AI.

**O1. Assessing State-of-the-Art of XAI in PM through Theoretical and Empirical Studies**

Conduct an extensive analysis on the current state-of-the-art in XAI for PM, encompassing literature review of techniques and barriers evaluation, aided also by empirical interviews and focus groups to discern limitations in existing technical approaches and interfaces.

**O2. Advancing SOTA Responsibly through Soft and Hard Laws**

Elucidate hard-law regulatory provisions and soft-law ethical considerations that must be accounted for in advancing XAI solutions beyond current technical capabilities in a responsible manner aligned with emerging policies, user needs, and organizational constraints.

**O3. Integrating SOTA and Soft/Hard Laws in a Conceptual Architecture for XAI-PM Implementation**

Develop a conceptual architecture encompassing explainability dimension, requirements elicitation, technical implementation, organizational integration, governance mechanisms and monitoring procedures to translate XAI innovations into pragmatic solutions fitted to real-world PM environments.

## 1.3 Research Contributions

The research developed in this Ph.D. dissertation find output in the contributions listed below:

1. **Systematic Literature Review of Explainable AI in PM**

   Beyond policies, we provide the first comprehensive literature review on the application of XAI techniques in PM. Through a systematic analysis of 45 studies, our findings reveal a heavy focus on predictive monitoring tasks using feature attribution methods like SHAP and LIME. In contrast, causal explanation methods remain underutilized across other PM tasks. Most studies rely on public benchmarks rather than real-world logs, restricting evaluation of practical utility. To address these gaps, we advance multi-layered recommendations encompassing cognitive, technical, organizational and pragmatic dimensions to advance XAI in PM. Our review establishes a baseline understanding to guide future research on integrating explainability into PM tools and practices.

2. **Client-Aligned Explainability Strategies in PM**

   Our two-phase study using a questionnaire and interviews reveals that current PM explainability approaches are often unstructured and context-specific. Key needs identified include structured frameworks, improved data integration, and assessing client maturity. Explainability effectiveness is influenced by delivery, barriers, contextuality, and maturity factors. The research highlights the need

for tailored strategies aligned with diverse stakeholders to empower responsible PM adoption, increasing organizational transparency. By elucidating dynamics, barriers, and improvements, the paper lays the groundwork for developing explainability strategies attuned to clients' needs and governance considerations.

3. **Evaluating AI Explainability Regulations – A Thematic and Gap Analysis**

   We presents the first comprehensive thematic and gap analysis of AI explainability policies and standards from the EU, US, and UK. Through rigorously surveying policy documents and contrasting them with research publications, the study identifies critical limitations in current explainability regulations. The analysis reveals that policies lack an informed perspective and provide discretion to providers without minimum requirements for explanations. Contributions include providing an overview of governmental regulatory trajectories within AI explainability and its sociotechnical impacts; mapping existing explainability regulations and standards across countries; conducting an intergovernmental gap analysis as informed by research constraints; analyzing discrepancies around explainability feasibility, usability, and accountability allocation; formulating evidence-based recommendations on defining and regulating explainability; and calling for policy documents cognizant of identified tensions to ensure responsible AI adoption.

4. **A Map of EU Regulations for AI Explainability – Constraints and Recommendations**

   We undertake a rigorous analysis of explainability requirements for AI systems across recent EU regulations and policies. Through a structured content analysis, it maps dimensions of explainability and involved stakeholders addressed in major EU legal documents. The mapping reveals tensions between desires for transparency to ensure oversight and accountability, versus legitimate secrecy around proprietary models and third party data. To balance these constraints, context-specific XAI approaches are recommended based on explanation targets, user needs, and deployment risks.

5. **Mapping Ethical Perspectives in XAI Research**

   We systematically analyze ethical considerations across XAI literature, revealing limited rigorous application of ethical theories. The study maps the prevalence of various ethical frameworks like consequentialism, deontology and

virtue ethics within XAI papers. It finds cursorily mentions of ethics without substantive analysis. To foster comprehensive integration, we recommend selecting appropriate ethical paradigms based on context, addressing inherent tensions, enabling interdisciplinary collaborations and educating developers. This analysis establishes a baseline to promote nuanced ethical discourse in XAI beyond ethics-washing.

6. **A Risk Management Framework for Ethics in XAI Implementation**

   We offer valuable perspectives on the integration of ethics within XAI research. Through systematic analysis, in a first review study we reveal a discrepancy between the frequent mentions of ethics in XAI papers and the limited rigorous application of ethical theories or frameworks. Many studies invoke ethics in a cursory, generic fashion without grounded ethical analysis. To address this, we set preliminary ethical strategies to deeply embed ethical considerations in XAI systems, from design to deployment. In this, we emphasize selecting appropriate ethical frameworks based on the context, proactively addressing tensions between principles, fostering interdisciplinary collaboration, and educating developers on ethical theories. Alongside that novel risk assessment framework is proposed to identify and mitigate technical and sociotechnical risks in XAI systems.

### 1.3.1 Publications

All the contributions shown in this dissertation are included in the following publications:

#### Journals

- Luca Nannini, Jose Maria Alonso-Moral, Alejandro Catala, Manuel Lama, and Senén Barro, **Operationalizing Explainable AI in the EU Regulatory Ecosystem**, *IEEE Intelligent Systems*, 2024. ISSN: 1541-1672. doi: 10.1109/MIS.2024.3383155.

  Journal Impact Factor (JCR 2023): *5.6*

  - 41/197 (Q1) *COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE*
  - 53/352 (Q1) *ENGINEERING, ELECTRICAL & ELECTRONIC*

- Luca Nannini, Alejandro Catala, Manuel Lama, and Senén Barro, **Explainable AI in Process Mining: A Systematic Literature Review**, *Springer GMBH Business & Information Systems Engineering*, Submitted. ISSN: 2363-7005.

  Journal Impact Factor (JCR 2023): *7.9*

  - 15/158 (Q1) *COMPUTER SCIENCE, INFORMATION SYSTEMS*

- Luca Nannini, Marta Marchiori Manerba, and Isacco Beretta, **Mapping the Landscape of Ethical Considerations in Explainable AI Research**, *Springer-Verlag Ethics and Information Technology*, ISSN: 1388-1957. doi: 10.1007/s10676-024-09773-7.

  Journal Impact Factor (JCR 2023): *3.6*

  - 8/57 (Q1) *ETHICS - SSCI*
  - 33/84 (Q2) *INFORMATION SCIENCE & LIBRARY SCIENCE - SSCI*
  - N/A *PHILOSOPHY - AHCI*

#### Conferences

- Luca Nannini, Agathe Balayn, and Adam Leon Smith. 2023. **Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK.** In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1198–1212. doi: 10.1145/3593013.3594074.

### 1.3.2 Participation in R&D projects, contract and networks

Aside from the publications reported, this Ph.D. thesis benefited and informed different projects:

- ***NL4XAI* – Natural Language for Explainable Artificial Intelligence**. This PhD project was made possible thanks to a European Union's training network where 11 *Early-stage Researcher* investigated dimensions of Exxplainable AI through natural language technologies. Through this network, this project benefited of a variety of training and secondments that sensitevly contributed to its outcome. Among those, two (2) deliverable technical reports were successfully

issued to the European Commission, namely *D4.1* and *D4.7* issued in 2020 and 2023.

- ***Minsait by Indra Sistemas* – Onesait products**. This project is an industrial doctorate developed at the facilities of a well-acquainted IT consulting company in Spain, *Minsait by Indra Sistemas*. This industrial PhD project benefited by this hosting institution, contributing sensitively to its knowledge through the management and implementation of language technologies. Such R&D regarded in particular the *Onesait Language* solutions, as well as successively the NLP solutions under the *Phygital* section. As of April 2024, the projects informs the AI Governance team in Minsait especially for EU regulations, with active consultancy to external clients.

- ***CAIDP* – Center for AI & Digital Policy**. This project was informed by a policy course taken during Fall 2022 by one of the most world famous organization in AI policy, the *Center for AI & Digital Policy* in Washington, DC, USA. The online semester course participation allowed to contribute to a final extensive annual report, the *CAIDP AI Policy Index 2022*. This thesis benefited from this contribution for the AI policy analysis.

- ***UNINFO - CEN-CENELEC & ISO-IEC* – AI Standard Activity**. The knowledge acquired during this PhD research allowed to transition to professional consultancy, done with the Italian Agency for Standards Setting *UNINFO* and the delegated participation *UNI/CT 533* within *CEN-CENELEC JTC21* AI Group (WG3), where I am actively working on defining standards in AI transparency and explainability as in accordance to the EU AI Act.

## 1.4 Methodology

The methodological framework employed in this dissertation is inherently multifaceted, adopting a mixed-methods paradigm with a predominant inclination towards qualitative research methodologies. This aligned with the complex, interdisciplinary nature of the research topic, necessitating diverse empirical lenses to develop a comprehensive understanding.

The initial phase involved an extensive systematic literature review, cataloging and analyzing technical publications on XAI techniques applied in PM contexts addressing objective O1. Quantitative bibliometric analysis coupled with qualitative thematic

synthesis enabled the methodical assessment of adoption patterns, influential contributions, and research gaps.

Building on this technical foundation, subsequent studies adopted qualitative methodologies better suited to capturing nuanced human perspectives regarding usability barriers, ethical tensions, and organizational constraints surrounding XAI-PM innovations. A combination of surveys, interviews, observational studies and analyses was conducted: expert focus groups shape responsible advancement proposals tackling regulatory and ethical complexities in objective O2; participatory explorations guide explanatory architecture development fitted for organizational constraints meeting objective O3.

## 1.5 Document Structure

The organization of this Ph.D. dissertation is designed to offer an in-depth and multifaceted analysis of XAI within the scope of PM.Figure 1.3 presents an overview of the overall research document structure based on the RQ defined earlier.

The remainder dissertation is structured as follows:

- *Chapter 2* provides a systematic literature review (SLR) for XAI approaches in PM to establish the theoretical foundations and identify gaps in the current research landscape.
- *Chapter 3* complements the SLR as a qualitative study with practitioners to gather empirical insights into real-world challenges and needs. It is constituted by a questionnaire and interviews, focusing on their strategies for enhancing explainability in process representations. With the SLR, these chapters map the current "state-of-the-art" (SOTA) of both theoretical and empirical approaches to enhance explainability in PM.
- *Chapter 4* conduct comparative analyses of international and EU policies related to AI explainability, drawing out themes around oversight explainability for authorities versus ambiguous individual rights. It discusses tensions in balancing transparency for accountability while protecting confidentiality.
- *Chapter 5* provides an ethical evaluation of XAI, serving as a normative guide for the technical and empirical discussions in subsequent chapters. In particular, the analysis of regulatory and ethical stances (Chapters 4 and 5) bridges the theoretical and empirical findings with organizational considerations (i.e.,

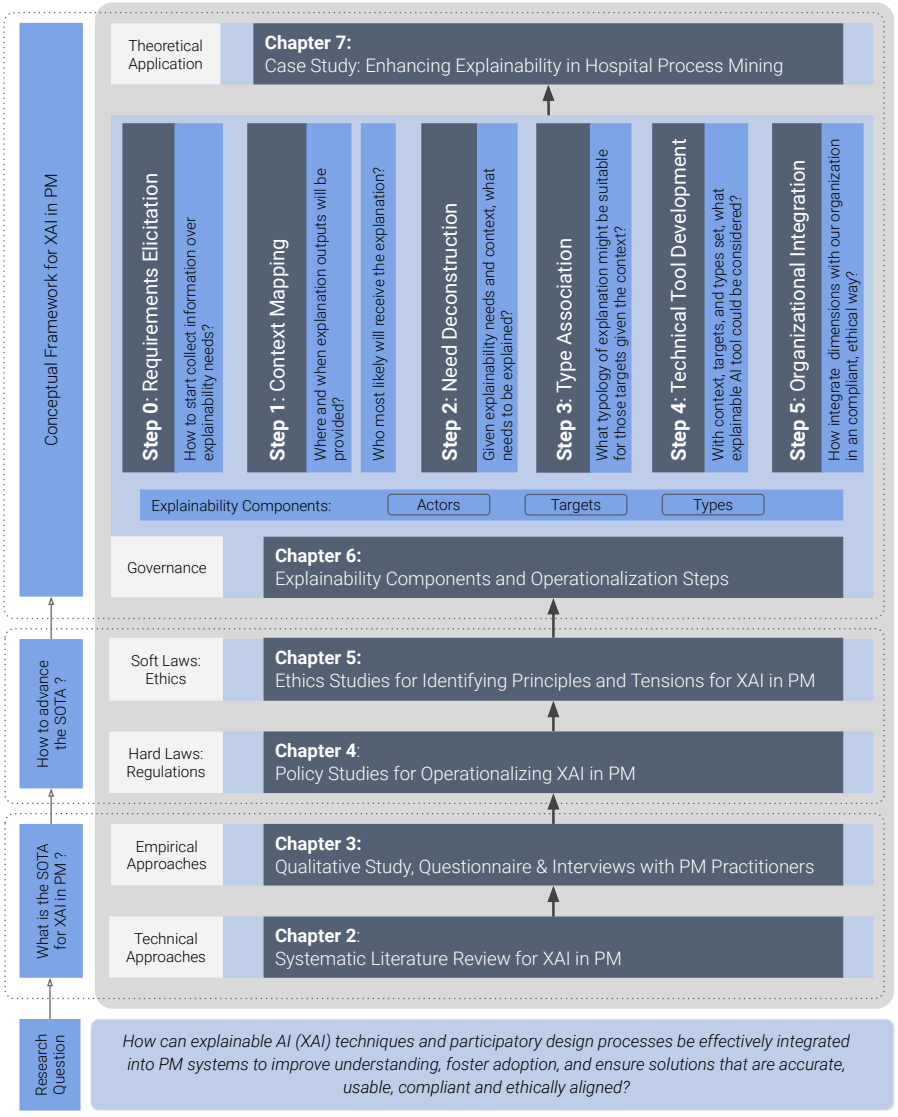


**Figure 1.3:** Bottom-up overview of the research structure and its contributions to advancing the state-of-the-art in algorithmic explainability for process mining.

advancing the SOTA) emphasizing the importance of aligning explainability approaches with both hard and soft laws.

- *Chapter 6* synthesizes insights from earlier chapters into a conceptual framework for operationalizing explainability in PM across several components, interactions components, and a proposed phased approach.
- *Chapter 7* discusses a theoretical application of the framework within the healthcare domain, drawn by a real use case with a Galician hospital network.
- *Chapter 8* summarizes the dissertation while addressing future research directions.
- *Appendices* provide additional material to supplement the empirical investigations conducted in Chapters 2 to 5, as well as Chapter 6 with Appendix C to aid self reflection when design explainable AI systems in process mining.

# Chapter 2

# Explainable AI in Process Mining: a Systematic Literature Review

In the introductory chapter, we provided necessary background on the key concepts of explainable AI (XAI) and process mining (PM), defining core terminology and outlining the motivation for explainability in PM. Building on that foundation, this chapter presents a systematic literature review examining the state-of-the-art in XAI techniques applied within PM. As an emerging interdisciplinary field leveraging AI methods like machine learning, PN is well-positioned to benefit from XAI techniques that can clarify model rationale and enrich process knowledge. However, research on XAI applications in PM remains in a nascent stage. A rigorous, structured analysis of the literature can reveal the scope of XAI techniques being applied, challenges faced, and gaps to be addressed. This chapter undertakes such a systematic review, scrutinizing the integration of XAI in PM research.

The chapter is structured as follows. First, related studies (Section 2.1) are exposed to detail the current research gap and the approach we intend to simulate. Then, the systematic search methodology is elaborated, detailing the strategy, screening process, data extraction, and trend analysis (Section 2.2). Next, the key results are presented (Section 2.4), and discussed to discern adoption patterns, influential works, and research gaps (Section 2.5). Finally, the limitations of the systematic literature review are acknowledged. By methodically assessing and structuring current knowledge, this review establishes a baseline understanding of the state, techniques, trends, and gaps in XAI-PM research.

## 2.1 Related Studies

The survey departs from detailing comprehensive review of related surveys in the fields of XAI and PM to situate it within the existing literature and highlight the research gap. Table 2.1 provides an overview of the key surveys, their focus areas, and limitations.

**Table 2.1:** Overview of related surveys on XAI and PM

| Survey | Focus Area | Limitations |
|---|---|---|
| [248] | Anomaly detection in process mining: classify techniques and identify future directions | • Limited focus on explainability<br>• Lack of labeled real-life logs |
| [119] | Process-aware recommender systems: identify recommendation types and dominant approaches & challenges | • Limited focus on interpretability<br>• Lack of standardized datasets and protocols |
| [419] | Concept drift in business processes: identify research branches while focusing on control-flow perspective and sudden drifts | • Lack of standardized datasets and metrics<br>• Overlooked explainability of models |
| [17] | XAI in Industry 4.0: identify applications of AI and XAI techniques and evaluation dimensions | • Barriers in model vulnerabilities and ethical alignment |
| [254] | Outliers and noise in event logs: categorization of outliers and noise type and analysis of detection algorithms | • Lack of emphasis on explainability of detected anomalies<br>• Limited understanding of effects on downstream tasks |
| [450] | Explanations in predictive business process monitoring: identify explanation purposes and methods, with a focus on post-hoc explanations for deep learning models | • Limited evaluation of explanations for end-users<br>• Lack of guidelines for measuring explanation quality |

Departing from the XAI domain, [17] explore the applications of AI and XAI techniques in Industry 4.0 across various domains. They discuss the use of machine learning, deep learning, and other AI methods in tasks like predictive maintenance and quality control, and examine different XAI methods. In their work, the authors highlight barriers in model vulnerabilities and adherence to moral codes. In the PM domain, [248] present a systematic review on anomaly detection in event logs. They classify existing techniques and identify future research directions, revealing a limited focus on explainability and the need for benchmark datasets. Similarly, [254] analyze 24 studies to categorize outliers and noise types in event logs and analyze detection algorithms, highlighting the lack of emphasis on the explainability of detected anomalies. [119] investigate 34 studies on process-aware recommender systems (PARS), identifying the main recommendation types and dominant approaches. However, they point out the limited focus on interpretability and the lack of standardized datasets and protocols. [419] review 45 studies on detecting and dealing with concept

drift in business processes, revealing a dominant focus on the control-flow perspective and sudden drifts. They emphasize the need for improved accuracy metrics, the integration of contextual data, and the development of explainable architectures. Focusing on predictive business process monitoring, [450] review 19 techniques that provide explanations to enhance the intelligibility of predictive business process monitoring models. They identify explanation purposes and discuss the shift towards post-hoc explanations for deep learning models, also highlighting limited evaluation of explanations for end-users and the lack of guidelines for measuring explanation quality.

These surveys collectively provide valuable insights into various aspects of XAI and PM. Yet they also reveal several limitations and gaps in the current research landscape, such as the lack of focus on explainability and interpretability, the need for standardized datasets and evaluation protocols, and the limited consideration of end-user perspectives and organizational factors in the adoption of XAI. Our SLR distinguishes itself from these existing works in several ways. First, while we share some classification criteria with [119], our SLR broadens the scope by examining application domains and the specific patterns and types of explanations within each domain. Second, we explicitly consider the explainability of AI techniques as a classification criterion applied broadly to all PM tasks, addressing a research gap identified by [254] and expanding over [450]. Finally, in addition to outlining open research directions, our SLR offers a research agenda to assist researchers in selecting and advancing suitable XAI methodologies for their specific application scenario within the context of explainable PM.

## 2.2 Methodology

## 2.3 Systematic Literature Review Methodology

To address specific Research Questions (RQs), this study conducts a systematic literature review (SLR) through a rigorous, multi-stage process [54]. Commencing with the definition of RQs in Section 2.3.1, the SLR unfolds in five distinct phases: first, the search strategy, elucidated in Section 2.3.2; second, the selection of relevant studies, expounded in Section 2.3.3, through their classification in a framework presented in 2.3.4; third, the analysis of the chosen studies, presented in Section 2.4; and finally, the synthesis and discussion of findings, covered in Section 2.5.

### 2.3.1 Research Goals & Questions

This review gathers, analyze and discuss evidence of specific RQs through literature pertaining to proposals for XAI within the realm of PM through the PRISMA Statement (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) [54,376]. Our goals are (i.) identify and categorize the landscape of XAI techniques that have been specifically tailored to extraction of insights from event data in PM applications, (ii.) understand tasks within PM where explainability has been deemed necessary by researchers and practitioners, (iii.) review their evaluation approaches, and (iv.) identify remaining open challenges and limitations. To reflect such goals, this SLR is based on the following key questions reported below in Table 2.2.

**Table 2.2:** Research Questions.

| | |
|---|---|
| 1. | **RQ1**: What methods have been proposed for explainable PM using AI techniques? |
| 2. | **RQ2**: For what PM tasks are these explanation techniques being applied? |
| 3. | **RQ3**: How are the explanations from AI techniques evaluated? |
| 4. | **RQ4**: What are the limitations and gaps in research on XAI for PM? |

### 2.3.2 Search Strategy

Departing from the four RQs, we define our search strategy for literature identification as follow. To develop a targeted yet wide-ranging search string, we adopted a "top-down" approach informed by them using terminological groups pertinent to RQ concepts like (G1) explainable AI techniques and (G2) PM tasks. We used an iterative strategy – starting with an initial search string, evaluating its performance against a test corpus of 15 relevant papers identified through informal channels, and expanding terms progressively. The final search strategy comprises a multifaceted string combining focused terms grouped under three key elements mapped to research questions:

- **G1 – Explainable AI Terminology:** The first component taps into a rich vocabulary around evolving XAI concepts like explainability, interpretability, transparency, understandability etc. as well as related techniques – both model-agnostic and model-specific. Variants of each term are covered for completeness. Overall, this facet corresponds to RQ1 on techniques. As references we

consulted prior surveys in the Explainable AI fields providing well-established taxonomy of XAI concepts – namely [5, 35, 73, 112, 329].

- **G2 – Process Mining Terminology:** To identify papers in the PM domain, the second major part focuses on process analytics tasks as well as input data concepts like event logs – corresponding to RQ2 on PM applications. The search terms cover the main PM tasks, including process discovery, conformance checking, performance analysis, and process enhancement [467]. Additionally, the search string includes terms related to event data, such as event logs and traces, which serve as the input for PM techniques [477]. To ensure comprehensive coverage, the search strategy also incorporates terms related to specific PM techniques, such as automated process discovery algorithms (e.g., Alpha algorithm, inductive mining, split miner) [467], conformance checking approaches [412], and predictive process monitoring methods [299]

The composite search string combines these three groups using Boolean AND operators, also harnessing the power of wildcard truncation (*) to account for variations: (G1) *AND* (G2)[1]. Database-appropriate variants are crafted to cater to specific syntax, functions and quirks of relevant scholarly platforms – primarly Scopus and ACM Digital Library, SpringerLink, Web of Science, and IEEE Xplore. Restrictions were limited to English language papers and timeframe of 2016 onwards – the year where the Explainable AI program from DARPA was officially announced [366].

To target PM publications, we considered specific PM-related conferences and journals, such as *International Conference on Business Process Management* (BPM), *International Conference on Advanced Information Systems Engineering* (CAiSE), *International Conference on Process Mining* (ICPM), *Transactions on Automation Science and Engineering* (IEEE), *Business Process Management Journal* (Emerald), and *Decision Support Systems* (Springer), *Information Systems* (Elsevier), *Business & Information Systems Engineering* (BISE - Springer), and *IEEE Transactions on Services Computing*. Apart from database searches, additional 'snowballing' search

[1]This search string strikes an optimal balance of wide coverage as well broad relevance to the RQs. The final complete search string deployed for the Scopus database (adjusted as required for other databases) is as follows: *("explainable artificial intelligence" OR "explainable AI" OR XAI OR explainability OR transparency OR understandability OR comprehensibility OR interpretability) OR ("interpretable machine learning" OR "interpretable ML" OR IML) OR ("explanation method*" OR "explanation model*" OR "explanation system*") AND ("process mining" OR "business process management" OR BPM OR "process model*") OR ("process discovery" OR "automated process discovery") OR ("conformance checking" OR "conformance analysis") OR ("process performance" OR "predictive process monitoring" OR PPM) OR ("process enhancement".) OR ( "log repair*" OR trace clustering").*

strategies was deployed for completeness– including manual screening of citations and references from included full texts. To manage search results systematically, records retrieved from across databases are exported to a Microsoft Excel Codebook and merged. The merged set then undergoes deduplication analysis within Excel to eliminate duplicates, comparing article titles, publication years, and digital object identifiers where available.

### 2.3.3 Study Selection

The study selection process aims to systematically filter papers identified through comprehensive search strategies to finalize the corpus of literature relevant to answering the research questions (RQ1-4). The process is defined with the following inclusion/exclusion criteria pertinent to the research scope, as outlined in Table 2.3.

**Table 2.3:** Inclusion and Exclusion Criteria.

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| 1. Published in peer-reviewed academic journals, conference proceedings, or books | Other formats: non-peer-reviewed articles, extended abstracts, tutorials, industrial white-papers, etc. |
| 2. Primarily deals with XAI and PM | Does not focus on XAI, PM, or related tasks and fields |
| 3. Provides novel methodologies, frameworks, or empirical findings for Explainable AI in PM | Mention focus terms, yet does not provide novel methodologies, frameworks, or empirical findings |
| 4. Written in English | Written in any other language |
| 5. Published after DARPA's XAI project (2016) | Published before 2016 |

Regarding the (2.) of the Inclusion Criteria, the threshold to determine inclusion is for papers that propose AI/ML methods for explainable PM (RQ1 relevance) i.e., the techniques are applied in any PM context (RQ2 relevance) e.g. discovery, conformance checking etc; and they involve evaluation of explanation quality (RQ3 relevance). The only exclusion criterion set pertains to reporting format - material like editorials, reviews, book chapters, once identified as pertinent by abstract content, will be excluded and instead hand searched for relevant cited literature. A cautionary approach was adopted with peer-reviewed journals considered 'predatory'[2].

The screening process was conducted by the main annotator, a Ph.D. candidate with expertise in XAI and PM. To mitigate the risk of bias during annotation, three professors (one with expertise in AI and PM, the other in XAI and HCI) cross-checked the extraction. In case of disagreements between the main annotator and

[2] I.e., we filtered out potential results from those venues due to difficulties from the academic community to validate their scientific contributions and publication policy.

the professors during the screening and quality check stages, the team discussed and reached a consensus, with the main annotator revising the cataloging if necessary.

The selection workflow consists of three stages:

1. **Stage 1 (Title/Abstract Screening):** The main annotator screens all records.

2. **Stage 2 (Full-Text Screening):** The main annotator screens all full-text articles. Three professors (two full professors in AI and PM, one assistant professor in XAI and HCI) cross-check subsets of the screened articles.

3. **Stage 3 (Quality Check):** The main annotator conducts a quality check. The professors review the coding for a randomly selected subset of articles.

In Stage 3 of the study selection process, we applied additional quality criteria (QC) to ensure that the selected studies were directly relevant to XAI in PM and to minimize the inclusion of false positives that did not truly engage with PM topics. These criteria, highlighted in Table 2.4 below, were designed to discriminate between studies that made substantial contributions to the field of XAI in PM and those that only superficially mentioned the relevant concepts.

**Table 2.4:** Quality Criteria for Screened Papers.

| ID | Quality Criteria (QC) |
|---|---|
| QC1 | Research objective clearly articulated |
| QC2 | Directly engaged with core PM concepts, terminology, or scholarly works |
| QC3 | Proposed methods to improve interpretability/ explainability of AI/ML techniques used for PM tasks |
| QC4 | Discussed tailoring explanation functionality for process analytics use cases |
| QC5 | Evaluated or analyzed explanation rigor |
| QC6 | Explicitly called out gaps, limitations, or constraints around explainable PM adoption |

Studies that met these quality criteria were included in the final analysis, while those that failed to satisfy them were excluded. Examples of excluded studies were those that e.g., focused exclusively on standard applications of AI in business processes without emphasis on explainability; dealt with predictive/prescriptive analytics without a process perspective; or mentioned explainability but did not propose new techniques or substantial insights tailored for PM tasks.

### 2.3.4 Classification Framework

To systematically analyze and categorize the selected studies on XAI in PM, we developed a comprehensive classification framework inspired by the approach used in recent systematic reviews on related topics, such as anomaly detection in PM [248] and explainable predictive business process monitoring [450].

Our framework consists of four main categories as reported in Figure 2.1: *XAI Techniques* (XAI-T), *Process Mining Tasks* (PMT), *Explanation Evaluation* (EE), and *Gaps and Limitations* (GL). These categories are designed to capture the key aspects of XAI in PM and align explicitely with our research questions.

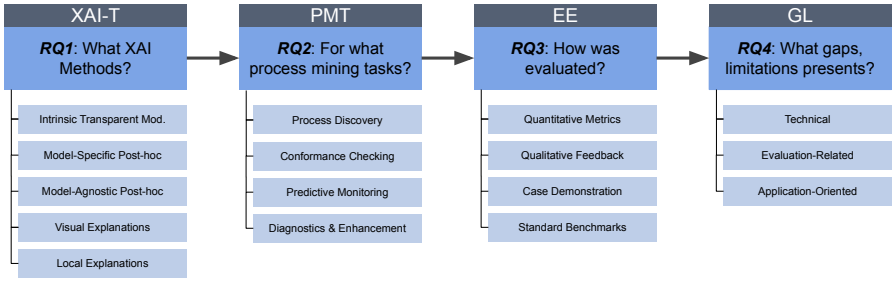


**Figure 2.1:** Classification Framework Reflecting the RQs.

1. The *XAI Techniques* (XAI-T) category aims to catalog the various XAI techniques implemented in the selected studies. We classify these techniques into five subcategories: *Intrinsic Transparent Models, Model-Specific Post-hoc, Model-Agnostic Post-hoc, Visual Explanations,* and *Local Explanations*. This categorization is based on established taxonomies and surveys on XAI techniques [5, 35], enabling a structured analysis of the methods employed in the PM context.

2. The *Process Mining Tasks* (PMT) category is included to understand the specific PM tasks addressed by the selected studies. PM encompasses a range of tasks, including *process discovery, conformance checking*, *predictive monitoring*, *process diagnostics and enhancement* among others, as also adopted in similar surveys in the field [119, 248, 254, 419].

3. *Explanation Evaluation* (EE) assesses the rigor of the explanation evaluation approaches employed in the selected studies. We consider quantitative metrics,

qualitative feedback, case demonstrations, and standard benchmarks as evaluation criteria. The inclusion of these criteria is justified by the growing emphasis on rigorous evaluation in the XAI literature [307, 329], which is essential for ensuring the reliability and practicality of XAI techniques in PM.

4. Finally, the *Gaps and Limitations* (GL) category captures any research gaps or adoption constraints discussed in the selected studies. Identifying these gaps and limitations is crucial for defining future research directions and informing practitioners about potential challenges in implementing XAI in PM.

We recorded metadata (i.e., bibliographic information such as authors, title, year, venue, etc) and then through such classification framework in a structured Excel spreadsheet we reported the content of each study, further nothing key insights and datasets (event logs) mentioned. For each study, we documented any missing or unclear information and attempted to resolve these issues by contacting the study authors via email. If no response was received within four weeks, we made conservative assumptions based on the available information and noted the limitations in the analysis.

## 2.4 Classification

In Figure 2.2, a flow diagram from PRISMA details the literature screening and selection, visaulizing the number of papers gathered, screened, and finally obtained. The PRISMA flow diagram consists of four main components:

1. **Identification**: The total number of records identified through the database and manual search, using the search strings and targeted conferences and journals mentioned earlier [=462];

2. **Screening**: The number of records remaining after duplicates have been removed, and the number of records excluded during the title and abstract screening based on the inclusion and exclusion criteria.

3. **Eligibility**: The number of full-text articles assessed for eligibility, and the number of articles excluded during the full-text review based on the inclusion and exclusion criteria.

4. **Included**: The final number of papers included in the study after the comprehensive two-phase screening procedure.

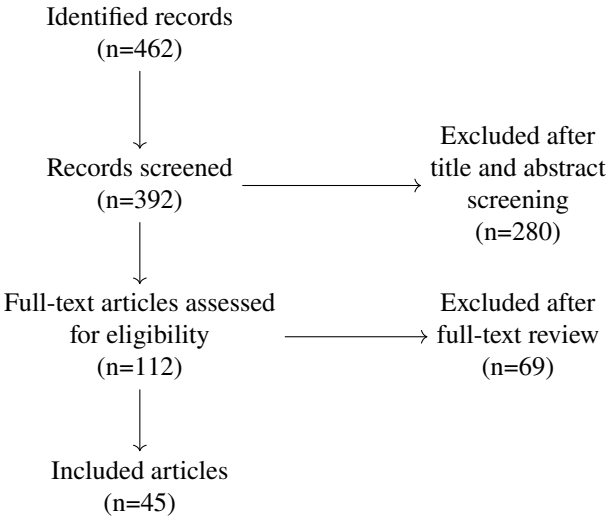


**Figure 2.2:** PRISMA Flow Diagram for Literature Screening Process.

### 2.4.1 RQ1: Methods for Explainable Process Mining

The included studies proposed a diverse set of methods for explainable PM, ranging from intrinsically transparent models to model-specific and model-agnostic post-hoc explanations, as well as visual and local explanations. These methods were applied to various PM tasks, with a strong focus on predictive monitoring. The choice of explanation technique often depended on the specific AI model being used and the desired level of interpretability. Table 2.5 summarizes key techniques and related studies.

#### Intrinsic Transparent Models

Intrinsic transparent models are inherently interpretable, allowing users to directly understand the decision-making process. In the context of explainable PM, three main types of intrinsic transparent models have been utilized: rule learners, linear models, and tree models.

- *Rule Learners* such as decision lists and RIPPER, have been employed to enhance the interpretability of PM techniques. For instance, [271] uses RIPPER to discover interpretable decision rules for each knock-out check in a business process, predicting if a case will be rejected. Similarly, [383] proposes FOX, a neuro-fuzzy model that learns a set of fuzzy IF-THEN rules representing the

| Techniques | Subcategory | Studies |
|---|---|---|
| **Intrinsic Transparent Models** | *Rule Learners* (Decision Lists, Ripper) | [271], [383], [500], [191] |
| | *Linear Models* (Regression, SVM) | [448], [428], [390] |
| | *Tree Models* (Decision Trees, Model Trees) | [428], [275], [448], [163], [207] |
| **Model-Specific Post-hoc Explanations** | *Attention Mechanisms* | [207], [498], [481], [448], [480], [143] |
| | *Layer-Wise Relevance Propagation* | [496] |
| **Model-Agnostic Post-hoc Explanations** | *Perturbation-Based* (LIME, Shapley, Anchors) | [122], [163], [379], [300], [278], [164], [275], [374], [448], [481], [480], [404], [432] |
| | *Gradient-Based* (Integrated SmoothGrad) | [106] |
| | *Counterfactuals* | [66], [208], [8], [339], [80], [206] |
| | *Prototype Selection* | [278], [163] |
| **Visual Explanations** | *Partial Dependency Plots* | [271], [390], [163], [164], [278] |
| | *Individual Conditional Expectations* | [504] |
| | *Saliency Maps* | [339], [169], [189], [496], [404], [143] |
| **Local Explanations** | *Instance-Level* | [481], [162], [379], [278], [163], [164], [382], [207], [80], [208], [404], [448], [7], [496] |
| | *Subgroup-Level* | [150], [163], [164], [80], [339], [428] |
| | *Log Slice-Level* | [496], [6], [169], [432] |

**Table 2.5:** Overview of explainable PM methods

predictive model for process outcome prediction. [500] utilizes rule learners in combination with minimum description length (MDL) principles to select compact, interpretable models for predicting event sequences from attributes. [191] employs fuzzy linguistic protoforms to automatically generate process descriptions, extracting temporal, frequency, and behavioral pattern indicators from discovered process models and event logs.

- *Linear Models* including regression and support vector machines (SVM), have been used to improve the explainability of PM methods. [448] employs logistic regression as an inherently interpretable model for outcome-oriented predictive process monitoring, comparing its explainability to post-hoc techniques like SHAP. [428] uses k-means clustering and linear models to discover context-aware process trees (CaTs) that capture data-based decisions and constraints, providing more explainable process models. [390] applies linear models like XGBoost and Random Forests on feature vectors including inter-case pattern features to develop interpretable models for remaining time prediction.

- *Tree-based Models* such as decision trees and model trees, have been widely adopted in explainable PM. [428] discovers context-aware process trees (CaTs) that capture data-based decisions and constraints, providing more explainable process models. [275] uses decision trees to generate trace saliency maps, highlighting significant activities for classifying process traces as simple or complex. [448] utilizes decision trees as an inherently interpretable model for outcome-oriented predictive process monitoring. [163] employs gradient boosted decision trees (Catboost) as an interpretable predictive model, comparing its explainability to black-box LSTMs using Shapley values. [207] uses Model-based Trees (MOB) which are inherently interpretable partition-based models for patient mortality risk prediction.

### Model-Specific Post-hoc Explanations

Model-specific post-hoc explanations aim to interpret black-box models by leveraging their internal structures or learned parameters. Two main approaches have been explored in the literature: attention mechanisms and layer-wise relevance propagation.

- *Attention Mechanisms* Attention mechanisms have been utilized to provide explanations for deep learning-based PM models. For example, [207] uses self-attention modules within the Vision Transformer architecture to intrinsically

highlight input feature relevance for predictions and extract attention maps for visualization. [498] proposes event and attribute attention to identify influential events and their attributes in process prediction tasks. [448] employs attention mechanisms for an LSTM model to explain outcome-oriented predictive process monitoring. [481] demonstrates an LSTM model using attention weights as an intrinsic explanation technique to identify feature importance for each prediction in a framework for inspecting predictive process monitoring models. [143] leverages attention weights from a convolutional neural network (CNN) to highlight input relevance aligned with predictions, using LIME to produce saliency maps of relevant regions in the input images for the CNN's predictions.

- *Layer-Wise Relevance Propagation* Layer-wise relevance propagation (LRP) has been employed to explain the predictions of deep learning models in PM. [496] introduces XNAP, which applies LRP to a trained Bi-LSTM model to determine the relevance of input activities for next activity prediction, visualizing the results as heatmaps.

### Model-Agnostic Post-hoc Explanations

Model-agnostic post-hoc explanations aim to interpret black-box models without relying on their internal structures or parameters. Four main approaches have been investigated in the literature: perturbation-based methods, gradient-based methods, counterfactuals, and prototype selection.

- *Perturbation-Based Methods* Perturbation-based methods, such as LIME (Local Interpretable Model-agnostic Explanations), Shapley values, and anchors, have been widely used in explainable PM. For instance, [122] evaluates global, model-agnostic post-hoc XAI methods, including SHAP, permutation importance, and accumulated local effects (ALE), to explain the overall behavior of predictive process monitoring models. [163] employs SHAP to quantify feature importance in an explainable decision support system for predictive process analytics. [379] uses SHAP values based on Shapley values from cooperative game theory to quantify each feature's contribution to the predicted decision by the ML model. [300] applies the model-agnostic, perturbation-based SP-LIME technique to explain monitoring model predictions, approximating the complex model locally using an interpretable model trained on perturbed copies of an instance. [278] uses SHAP with PartitionExplainer for BERT text models and TreeExplainer for XGBoost tabular models to explain predictions in

a document-enriched predictive process monitoring framework. Several other studies [164, 275, 374, 404, 432, 448, 480, 481] also employ perturbation-based methods like LIME and SHAP to explain predictions in various PM tasks.

- *Gradient-Based Methods* Gradient-based methods, such as integrated gradients and SmoothGrad, have been applied to provide explanations for PM models. However, their usage has been limited compared to perturbation-based methods. [106] uses integrated gradients, a perturbation-based approach, in combination with the Alibi Explain library to provide instance-level explanations for a deep learning model predicting process failures.

- *Counterfactuals* Counterfactual explanations have gained attention in explainable PM. [66] proposes an evaluation framework to compare different approaches for generating diverse counterfactual explanations in outcome-based predictive process monitoring. [208] introduces LORELEY, a counterfactual explanation technique tailored for predictive business process monitoring tasks, extending the LORE technique to handle multi-class prediction problems. [8] uses Granger causality to uncover counterfactual scenarios in an explainable concept drift detection framework. [339] combines reinforcement learning with causal discovery and reasoning to identify causal relationships between events and enable reasoning about alternative action paths. [80] generates explanations for anomalies detected in business process event logs using counterfactual reasoning, comparing an anomaly with its most similar normal counterparts to identify unusual behaviors. [206] proposes DiCE4EL, a counterfactual explanation algorithm extending the popular DiCE technique, specifically tailored for event log data, using a new loss function combining class loss, distance loss, category loss, and scenario loss to handle categorical variables and ensure counterfactual validity as per the process model.

- *Prototype Selection* Prototype selection methods have been used to provide explanations in PM. For example, [278] employs SHAP with PartitionExplainer for BERT text models and TreeExplainer for XGBoost tabular models to explain predictions in a document-enriched predictive process monitoring framework. The prototype visualizes feature importance for individual cases and globally. [163] uses SHAP as a prototype selection method to quantify feature importance in an explainable decision support system for predictive process analytics.

### Visual Explanations

Visual explanations aim to present the decision-making process of PM models in an intuitive and interpretable manner. Three main approaches have been explored in the literature: partial dependency plots, individual conditional expectations, and saliency maps.

- *Partial Dependency Plots* Partial dependency plots have been used to visualize the relationships between input features and model predictions in PM. For instance, [271] visualizes the distributions of numerical case attributes in decision rules, highlighting the values of rejected cases to aid rule adjustment. [390] employs performance spectra plots to visualize case performance and error progression within process segments, enabling visual analysis to detect uncertain segments susceptible to inter-case dynamics. [163] visualizes feature importance as bar charts showing average Shapley value magnitude per feature, indicating positive or negative influence on predicted KPIs. [164] visualizes feature importance as boxplots showing Shapley value distribution, indicating positive/negative influence on predicted outcomes. [278] visualizes feature importance for individual cases and globally using a prototype based on SHAP values.

- *Individual Conditional Expectations* Individual conditional expectations (ICE) plots have been utilized to provide personalized explanations for PM models. [504] uses SHAP values and ICE plots to explain feature importance in predicting recurrence for patients with ischemic cerebrovascular events based on process discovery and transfer learning.

- *Saliency Maps* Saliency maps have been widely adopted to highlight the importance of input features or model components in PM explanations. For example, [339] uses edit distance-based case similarity as a form of saliency map, assigning higher weights to normal cases more similar to an anomaly, and utilizing these similar cases to generate anomaly explanations. [6] uses saliency maps to visualize SHAP values for individual predictions in an object-centric PM framework. [169] infers a multi-perspective likelihood graph from a next event predictor, employing exhaustive case generation and threshold-based filtering as a saliency map to highlight important process behavior. [189] generates Directly Follows Graphs (DFGs) as saliency maps highlighting transition probabilities learned by an LSTM model for next activity prediction. [496] vi-

sualizes the relevance values obtained from layer-wise relevance propagation (LRP) as heatmaps, highlighting the salient activities and their contribution towards the predicted next activity for a given prefix trace. [404] uses LIME to provide saliency maps highlighting important features for process outcome predictions. [143] produces saliency maps by highlighting relevant regions in input images for a CNN's predictions using LIME.

### Local Explanations

Local explanations focus on providing interpretations for specific instances, subgroups, or slices of the process data. Three main levels of local explanations have been considered in the literature: instance-level, subgroup-level, and log slice-level.

- *Instance-Level Explanations* Instance-level explanations aim to provide interpretations for individual process instances. For example, [481] proposes a framework to guide the generation of model explanations for the purpose of inspecting process predictive models, considering explanations at different granularity levels, including instance-level explanations for individual predictions. [162] generates local explanations for each running case, indicating the key factors driving its prediction. [379] provides local explanations for each running case to explain the predicted decision for that specific instance. [278] shows SHAP values indicating each feature's impact per case in a prototype for document-enriched predictive process monitoring. [163] computes Shapley values to quantify feature importance for individual running cases, displayed as a bar chart of most influential features. In another work, those authors [164] also quantifies feature importance for individual object instances using Shapley values. [80] generates linguistic summaries for each anomalous case individually by comparing it to its specific set of similar normal cases, thereby providing local, instance-level explanations. [208] provides local explanations for individual process instances, showing minimal changes to achieve a desired outcome. [404] derives explanations for individual trace predictions using LIME. [448] generates explanations per test prefix for outcome-oriented predictive process monitoring. [7] visualizes relevance scores for exemplary loan application instances along with accept/reject predictions, identifying key activities influencing the predictions. [496] determines relevance values for input activities by backward propagating the prediction through the network layers

using specific propagation rules for weighted connections and multiplicative interactions, providing instance-level explanations.

- *Subgroup-Level Explanations* Subgroup-level explanations focus on providing interpretations for specific subgroups of process instances sharing common characteristics. For instance, [150] generates descriptions relevant for traces/-cases with specific attributes in the Process-To-Text (P2T) framework. [163,164] aggregate instance-level local explanations to provide subgroup-level explanations, showing the overall impact of features on predictions across cases with common attribute values. [80] provides explanations for subgroups of cases with similar anomalies by identifying frequent deviating patterns. [339] represents event patterns explaining behaviors for KPI-based subgroups in discovered Petri net transitions. [428] defines contexts shared by trace subsets and provides constraints scoped to subtrees in context-aware process trees (CaTs).

- *Log Slice-Level Explanations* Log slice-level explanations aim to provide interpretations for specific slices or segments of the process event log. For example, [496] demonstrates the explainability of XNAP by showing directly follows graphs (DFGs) for specific process instances, providing explanations at the log slice level, while [189] mentions that DFGs provide explanations at the global log level. [26] suggests that model-informed LIME can enable explanations at the level of slices/segments of the event log. [6] infers a multi-perspective likelihood graph from a next event predictor, explanaining a slice of the log deemed likely by the predictor. [169] provides local explanations at the subgroup-level and log slice-level by representing different variants/paths through the process in the inferred likelihood graph. [432] generates feature importance explanations for different trace length buckets in predictive monitoring.

### 2.4.2 RQ2: Process Mining Tasks and XAI Techniques

The included studies applied XAI techniques to a range of PM tasks, with a strong focus on predictive process monitoring. The choice of XAI technique often depended on the specific sub-task and the underlying AI model being used, with interpretable models, post-hoc explanation methods, and visual explanations being the most common approaches. Figure 2.3 represents the numbers of identified PM tasks with respect to XAI techniques, while at the end of the Chapter an expanded Table 2.14 with references is reported.


**Figure 2.3:** Identified process mining tasks frequency with respect to XAI Methods.

### Predictive Process Monitoring

Predictive monitoring is the most extensively studied PM task in the context of XAI, with numerous papers proposing various approaches to predict process outcomes, next activities, and performance indicators while providing interpretable explanations.

- *Outcome Prediction* Several papers focus on binary outcome prediction, such as classifying process instances as successful or unsuccessful. [163] propose an explainable predictive monitoring framework using SHAP to explain the acceptance or rejection of loan applications. Similarly, [382] employ SHAP to explain customer churn predictions in a retail setting. [207] introduce a transparent sequence model based on event-flow graphs to predict patient mortality risk. Other papers address outcome prediction in various domains, such as hospital processes [275], manufacturing [449], and software development [191].

- *Next Acitivity Prediction* is another prominent focus, where the objective is to forecast the most likely next event in a running case. [498] propose attention-based neural networks to predict next activities while explaining the influential event attributes. [496] employ layer-wise relevance propagation to generate explanations for LSTM-based next activity predictions. [191] utilize gated graph neural networks to predict next activities and visualize the explanations as process models. Other papers explore next activity prediction using counterfactual analysis [206], decay replay mining [457], and pattern matching [500].

- *Performance prediction*, especially remaining time estimation, is addressed in several papers. [482] propose a white-box approach based on process models to provide explainable remaining time predictions. [53] introduce a technique called *LoGo* that combines local and global models for remaining time prediction while preserving interpretability. [379] employ explainable predictive process monitoring to estimate remaining time, next activity occurrence, and case cost. Other papers focus on predicting case duration [163], cycle time [164], and process performance indicators [340].

- *Concept drift* and *inter-case dynamics* pose additional challenges in predictive monitoring. [8] propose a framework for explainable concept drift detection, using time series analysis and causal reasoning to uncover the root causes of performance changes over time. [382] develop an online learning approach called TSUNAMI to adapt to concept drifts in customer churn prediction while explaining the importance of input features. [390] address inter-case dynamics in remaining time prediction by identifying high-variance process states and incorporating inter-case features into predictive models.

Other papers explore various aspects of explainable predictive monitoring, such as the impact of textual data [278], stability of post-hoc explanations [481], and adversarial attacks [449]. Some papers focus on specific application domains, such as healthcare [100, 275, 504], manufacturing [106, 449], and public administration [481].

| Subcategory | Studies |
|---|---|
| **Outcome Prediction** | [53, 162, 163, 191, 207, 275, 382, 383, 398, 404, 449, 482] |
| Next Activity Prediction | [169, 189, 191, 206, 381, 448, 457, 496, 498, 500] |
| Performance Prediction | [53, 163, 164, 339, 340, 374, 379, 390, 482] |
| Remaining Time Prediction | [53, 374, 379, 390, 482] |
| Concept Drift and Inter-case Dynamics | [8, 382, 390] |
| Explanations for Predictive Monitoring | [100, 106, 163, 164, 206, 275, 278, 313, 432, 448, 449, 480, 481, 499, 504] |

**Table 2.6:** Subcategories of predictive process monitoring

### Process Discovery

Table 2.7 outlines the role of explainable AI in process discovery, where techniques aim to extract interpretable process models from event logs.

[7] propose a framework for explainable concept drift detection, combining time series analysis and causal reasoning to identify root causes of changes in process behavior over time. By transforming event logs into time series for different perspectives, detecting change points, and performing causal analysis, the framework uncovers relationships between perspectives. [207] introduce an interpretable model-based tree Markov model for discovering process models from event sequences. Model-based trees partition data into subgroups with distinct patterns, while hidden semi-Markov models capture temporal dynamics within each subgroup. Visual explanations play a crucial role in making discovered process models more comprehensible. [189] generate directly follows graphs from LSTM predictions to visualize learned process behavior, highlighting probable activity transitions. [208] convert Petri net reachability graphs into Markov decision processes, enabling visual reasoning about event causality and alternative process paths.

Other works incorporate additional perspectives beyond control flow for interpretability. [428] integrate data-based decisions and constraints into discovered models, capturing the interplay between control and data flow. [142] introduces a specification-based approach to extract relevant process details from logs and annotate existing models, enhancing completeness and interpretability. [150] propose the Process-To-Text (P2T) framework, which integrates PM with natural language generation (NLG) techniques to automatically provide textual descriptions of discovered process models. The framework extracts temporal, frequency, and behavioral patterns from logs and generates human-readable explanations using hybrid template-based NLG.

| **Subcategory** | **Studies** |
|---|---|
| Explainable Concept Drift Detection | [7, 8] |
| Interpretable Process Model Discovery | [142, 207, 428] |
| Visual Explanations for Discovered Models | [189, 208] |
| Natural Language Explanations for Process Models | [150] |

**Table 2.7:** Subcategories of process discovery

### Conformance Checking

Table 2.8 summarizes the subcategories of conformance checking addressed by the included studies. These studies employ XAI techniques to provide interpretable insights into the alignment between observed process behavior and reference process models.

[334] propose a multi-perspective conformance checking approach that incor-

porates control flow, data, and resource constraints into the alignment computation, enhancing the explainability of the resulting alignments. The deviating patterns identified provide explanations of the anomalies in the contexts in which they occur. Entropic relevance is introduced by [21] as an information-theoretic measure for stochastic conformance checking, providing an interpretable measure of how well the model describes the observed behavior. Visual explanations, such as linguistic summarization based on fuzzy logic [80] and trace saliency maps [275], are employed to make conformance checking results more interpretable. These approaches generate human-readable explanations of detected anomalies and highlight the most significant activities influencing the classification decision. [383] with the neuro-fuzzy model FOX predict and explain process outcome by learning interpretable fuzzy rules from event log data.

| **Subcategory** | **Studies** |
|---|---|
| Multi-perspective Conformance Checking | [334] |
| Stochastic Conformance Checking | [21] |
| Visual Explanations for Conformance Results | [80, 275] |
| Anomaly Detection and Explanation | [383] |

**Table 2.8:** Subcategories of conformance checking

### Diagnostics and Enhancement

Table 2.9 summarizes the applications of XAI in diagnostics and process enhancement tasks. This subsection explores how these techniques aid in identifying performance bottlenecks, explaining root causes, and suggesting improvements to optimize processes.

[80] propose a model-agnostic method for explaining anomalies in event logs. By comparing anomalous cases with similar normal ones, their approach highlights differences in activity sequences and throughput times using linguistic summarization and quantified propositions. For complex industrial processes, [340] presents a data-driven causality analysis approach. The discovered dynamic causal models, represented as Petri nets, provide visual representations of system dynamics, enabling root cause identification for performance issues. Addressing overprocessing waste, [271] focuses on discovering and mitigating knock-out checks through interpretable decision rules and visual attribute distributions, suggesting redesign options like reordering or adjusting decision rules. Other works explore domain-specific applications, such as [137]'s causal process discovery for explaining delays in pizza

delivery, using causal narratives as input to large language models. [404] investigate the trade-off between explainability and performance in predictive process monitoring, while [398] demonstrate the potential of interpretable models for process optimization and human-machine collaboration in an Industry 4.0 context.

| Subcategory | Studies |
|---|---|
| Anomaly Explanation | [80] |
| Causal Analysis for Root Cause Identification | [137, 340] |
| Overprocessing Waste Discovery and Redesign | [271] |
| Explainable Process Improvement | [339, 374, 398, 404] |

**Table 2.9:** Subcategories of diagnostics and process enhancement

### 2.4.3 RQ3: Evaluation of Explanations

Evaluating the quality and effectiveness of explanations is a critical aspect of XAI in PM. The included studies employed a variety of evaluation approaches, ranging from quantitative metrics to qualitative feedback, case demonstrations, and standard benchmarks. Table 2.10 showcase valuable evaluation approaches employed in the surveyed papers.

#### Quantitative Metrics

Several studies introduced novel quantitative metrics to assess the quality of explanations. [122] proposed consistency metrics (Consistency Ratio, AICConsistency, and BICConsistency) to quantify the alignment between feature attribution explanations and principal data features. Prediction quality metrics like accuracy, F1-score, and MAE were commonly used to evaluate the performance of explainable models (e.g., in [163, 381, 383]). Some studies proposed metrics tailored for specific explanation techniques. [66] introduced quality metrics (distance, implausibility, diversity, avg.changes), hit rate, runtime, and a process constraint compliance score to extensively evaluate counterfactual explanations. [207] applied AUROC to assess the classification performance of different transparent models on imbalanced datasets. [21] proposed entropic relevance as an information-theoretic measure of model quality for stochastic conformance checking. [448] introduced parsimony, functional complexity, and monotonicity to quantify the explainability of predictive monitoring models. Other quantitative evaluations focused on the faithfulness of explanations to the underlying models. [26] measured the percentage of perturbed instances adhering to process constraints in their extended LIME approach, while [480] proposed metrics to assess

the stability and fidelity of explanations generated by LIME and also SHAP. [208] used a fidelity metric to assess how well the local interpretable model approximates the black-box predictions in their counterfactual approach LORELEY.

### Qualitative Feedback

Qualitative feedback from domain experts was recognized as a valuable approach to assess the understandability and usefulness of explanations. [163] carried out a study with 20 process analysts to evaluate the perceived difficulty and usability of explanations through questionnaires. [26] had two independent raters with business process modeling experience assess the adequacy of LIME explanations extended with process constraints. [143] mentioned gathering feedback on the interpretability of process details and their representation in enhanced process models. [6] suggested using expert feedback to evaluate explanation quality for data-driven models in metal forming processes. [339] emphasized the role of process experts in validating insights on event causality, KPI impacts, and corrective actions derived from their causal models.

Some studies proposed metrics tailored for specific explanation techniques. [66] introduced quality metrics (distance, implausibility, diversity, avg.changes), hit rate, runtime, and a process constraint compliance score to extensively evaluate counterfactual explanations. [207] applied AUROC to assess the classification performance of different transparent models on imbalanced datasets. [21] proposed entropic relevance as an information-theoretic measure of model quality for stochastic conformance checking. [448] introduced parsimony, functional complexity, and monotonicity to quantify the explainability of predictive monitoring models. Other quantitative evaluations focused on the faithfulness of explanations to the underlying models. [26] measured the percentage of perturbed instances adhering to process constraints in their extended LIME approach.

### Case Demonstration

Demonstrating the proposed explainable techniques on real-world case studies was a common evaluation approach. This allowed showcasing the practical applicability and value of the methods. Case studies spanned various domains such as healthcare [150, 383], insurance claims [278], manufacturing [143, 398], and loan applications [191, 208]. Detailed demonstrations were provided to illustrate how the explanations offer meaningful insights into process behavior. [137] demonstrated their

causal process discovery approach on a simulated pizza delivery process to show how the causal narrative enables precise explanations of delays. [339] validated their causal reinforcement learning approach on two real-world heat recovery case studies to explain energy savings.

### Standard Benchmarks

Several studies utilized standard benchmark datasets to facilitate comparative evaluation and reproducibility. The Business Process Intelligence Challenge (BPIC) logs were widely used, including BPIC 2011 [404], BPIC 2012 [163, 383], BPIC 2013 [163], and BPIC 2017 [491]. These real-life logs from various domains served as common testbeds for evaluating predictive performance and explanation quality. Other benchmarks included the Sepsis Cases log [122, 383], Road Traffic Fine Management log [21, 163], and a synthetic log for assessing deviance mining [94]. By testing on multiple benchmark datasets, the robustness and generalizability of the explainable techniques could be demonstrated.

**Table 2.10:** Summary of explanation evaluation approaches in the surveyed papers

| Evaluation Approach | Papers |
|---|---|
| Quantitative Metrics | [21, 26, 66, 122, 163, 207, 208, 381, 383, 448] |
| Qualitative Feedback | [26, 143, 163, 339] |
| Case Demonstration | [137, 143, 150, 191, 208, 278, 339, 383] |
| Standard Benchmarks | [21, 94, 122, 163, 383, 404, 491] |

## 2.4.4 RQ4: Limitations of Current Research

Several limitations arisen from the paper pool. These limitations can be categorized into three main areas: technical, evaluation-related, and application-oriented. Table 2.11 summarizes the main limitations identified.

### Technical Limitations

Current XAI techniques for PM face *scalability and generalizability* challenges. Many methods are evaluated on small or synthetic datasets, and their performance on large, real-world logs is unexplored. Techniques like counterfactual explanations [66, 208], causal modeling [137], and natural language explanations [150] have shown promise but need further investigation in diverse PM contexts. Conformance checking, in particular, has received less attention compared to predictive monitoring and discovery tasks [21, 334]. In this vein, a similar technical limitation is the *lack of consideration*

**Table 2.11:** Limitations of current research on explainable AI in process mining.

| Category | Limitations |
|---|---|
| *Technical* | - Scalability and generalizability of XAI techniques to large-scale, real-world process logs<br>- Limited applicability to different PM algorithms and data formats<br>- Lack of consideration for temporal and dynamic aspects of PM data |
| *Evaluation-related* | - Lack of standardized benchmarks and evaluation metrics<br>- Heterogeneity in evaluation methods, hindering comparability across studies<br>- Limited user-centric evaluations considering cognitive and domain-specific factors |
| *Application-oriented* | - Limited exploration of XAI in conformance checking and process enhancement tasks<br>- Underexplored XAI application in specific domains (e.g., healthcare, manufacturing)<br>- Lack of domain-specific research considering unique challenges and requirements |

*for the temporal and dynamic aspects of PM data* [207, 382]. Process logs often contain time-related information, such as event timestamps and durations, which can be crucial for understanding and explaining process behavior. Several studies have also highlighted the challenge of generating explanations that are both *faithful to the underlying models and comprehensible to users*. Techniques like LIME and SHAP provide local approximations but may not fully capture the complexity of the original models [191, 404]. More research is needed on explanation methods that provide a good balance between fidelity and interpretability, as well as techniques for considering the causal relationships and limitations of the explanations [62, 448].

### Evaluation-Related Limitations

The analyzed paper pool denoted a lack of standardized benchmarks and evaluation metrics for assessing explanation quality and effectiveness in PM. While studies propose metrics, user studies, and case studies, a comprehensive framework combining these approaches rigorously is missing, thus limiting the practical value of the generated explanations. *Establishing benchmarks and evaluation tasks specifically for explainable PM* would enable comparative analysis [7, 66, 122]. Deeper *engagement with domain experts* is needed to understand their requirements and evaluate usefulness in real-world decision contexts. Collaboration across PM, AI, and industry is crucial, yet user-centric evaluations considering cognitive and domain factors influencing explanation effectiveness are scarce [163, 404].

### Application-Oriented Limitations

There are several application-oriented limitations in the current research on XAI for PM. One major limitation is the *limited exploration of explainable techniques for certain PM tasks, such as conformance checking*. Few studies have addressed the challenges of providing meaningful explanations for deviations between observed process behavior and a reference model [21, 334], despite the importance of conformance checking for identifying non-compliant or inefficient processes. Similarly, the application of XAI in process enhancement and optimization tasks remains underexplored [137, 339]. Explainable techniques that can provide insights into these tasks could greatly support data-driven process optimization efforts, but current research in this area is limited.

Zooming out, *the integration of XAI techniques into end-to-end PM pipelines and tools is also an important consideration not often taken into account*. While some studies have proposed prototype implementations [142, 163], more work is needed on developing user-friendly interfaces and workflows that enable seamless interpretation and interaction with PM results [404].

## 2.5 SLR Synthesis and Discussion

A total of 45 studies were included in this qualitative synthesis, distributed as 15 Journal Papers and 30 Conference Proceedings Papers. As XAI is relatively new research area gauging attention in the last fee years, the selected studies reflect this trend, as reported in Figure 2.4.

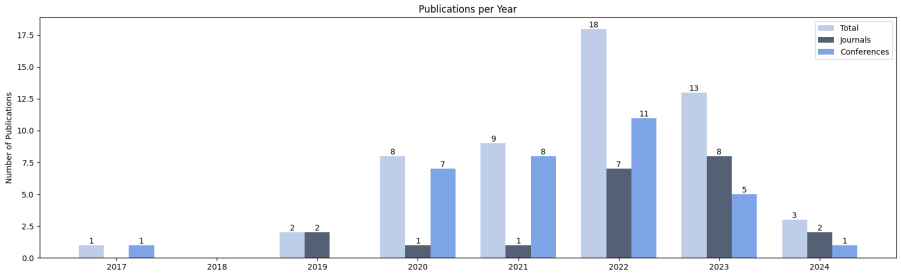


**Figure 2.4:** Frequency of studies per publication year

The studies covered a range of PM tasks, with predictive monitoring being the most common (38 studies), followed by diagnostics (9 studies), process discovery (8 studies), enhancement (8 studies), and conformance checking (6 studies). The studies

also employed a diverse set of XAI techniques, with model-agnostic post-hoc methods being the most prevalent (31 studies), particularly perturbation-based techniques like LIME, SHAP, and counterfactuals (21 studies). Intrinsic transparent models, such as rule learners, linear models, and tree models, were used in 18 studies. Visual explanations, including saliency maps and partial dependency plots, were employed in 16 studies. Model-specific post-hoc techniques, such as attention mechanisms and layer-wise relevance propagation, were used in 8 studies. Local explanations at the instance-level, subgroup-level, or log slice-level were provided in 14 studies.

Figure 2.12 present a summar of publication venues. The most frequent were the *International Conference on PM* (ICPM) with 6 papers, followed by the *International Conference on Business Process Management* (BPM) with 4 papers, and the I*nternational Conference on Advanced Information Systems Engineering* (CAiSE) with 4 papers. Other notable venues included the *International Conference on Service-Oriented Computing* (ICSOC) with 3 papers and the *International Conference on Enterprise Information Systems* (ICEIS) with 3 papers.

| ***Journal Sources* = 15** | ***Conference Sources* = 30** |
|---|---|
| • Information Systems (4) | • CAiSE (4) |
| • Eng. Appl. Artif. Intell. (2) | • ICPM (6) |
| • Expert Syst. Appl. (1) | • BPM (4) |
| • J Intell Inf Syst (1) | • ICSOC (4) |
| • IEEE J. Biomed. Health Informatics (1) | • ICEIS (3) |
| • IEEE Transactions on Art. Intell. (1) | • CAiSE Forum (2) |
| • Knowl. Based Syst. (1) | • FUZZ-IEEE (1) |
| • Decis. Support Syst. (1) | • IJCAI (1) |
| • J. Biomed. Informatics (1) | • EMCIS (1) |
| • PLOS ONE (1) | • RCIS (1) |
| • J. Softw. Evol. Process (1) | • MIPR (1) |
| • IEEE Access (1) | • TAILOR (1) |
| • Künstliche Intell. (1) | • ICSOC Workshops (1) |
| | • IJCAI Workshops (1) |

**Table 2.12:** Publication Venues

The rigor of explanation evaluation varied across studies, with case demonstrations being the most common approach (32 studies), followed by quantitative metrics (24 studies), qualitative feedback (6 studies), and standard benchmarks (6 studies). Key research gaps and limitations identified include the need for more rigorous explanation quality measures, user studies to validate explanation usefulness, and extending the techniques to handle complex real-world process characteristics like concurrency, noise, and drift.

The most frequently used datasets for evaluation were real-world event logs from

the B*usiness Process Intelligence Challenge* (BPIC), with 19 studies using BPIC logs from various years (2011, 2012, 2013, 2015, 2017, 2019, 2020). Other commonly used real-world logs included the *Sepsis Cases* log (4 studies), the *Road Traffic Fine Management* log (3 studies), and the *Helpdesk* log (3 studies). Additionally, 12 studies used synthetic or self-collected datasets to demonstrate their approaches.

### 2.5.1 Implications of Findings

Our systematic review observed a diverse range of XAI methods applied to various PM tasks, with a strong emphasis on predictive monitoring. The proposed techniques encompass intrinsic transparent models, model-specific and model-agnostic post-hoc explanations, as well as visual and local explanations [164, 381, 404, 496]. Moreover, our findings underscore the importance of considering the temporal and dynamic nature of PM data when developing explainable techniques [207, 382]. This calls for exploring methods that can effectively capture and explain complex relationships and dependencies within event logs, such as causal modeling [137] and reinforcement learning [339].

The need for rigorous evaluation protocols and standardized benchmarks to assess explanation quality and effectiveness is evident from our findings [66, 122]. Furthermore, the limited focus on user-centric evaluations [404] highlights the importance of considering cognitive and domain-specific factors influencing interpretability and usability. The application-oriented limitations identified suggest a need for targeted research in underexplored areas like explainable conformance checking [21, 334], process enhancement [137, 339], and domain-specific applications e.g., healthcare [150] or manufacturing [398].

### 2.5.2 Addressing Limitations

To overcome technical limitations, we propose:

1. Developing distributed and parallel computing frameworks leveraging advanced data processing and storage for handling large-scale, real-world logs [300].

2. Creating adaptive and flexible XAI techniques applicable to different algorithms and data formats using modular architectures [163, 432].

3. Developing time-aware and event-based XAI techniques capturing the evolving nature of processes using temporal logic or event calculus [207, 382].

4. Implementing robust and efficient XAI techniques handling complexity and variability of real-world logs using optimization and user-friendly interfaces [143, 208].

To enhance evaluation practices, we propose developing standardized benchmarks comprising diverse real-world and synthetic logs spanning different domains, process characteristics, and requirements [66,122]. These benchmarks should include ground-truth explanations for evaluating accuracy and comprehensibility. Holistic evaluation frameworks integrating quantitative and qualitative methods are recommended to assess fidelity, interpretability, and actionability [26, 163]. User studies, interviews, and surveys can gather feedback on usefulness and usability from experts and end-users. Indeed, incorporating cognitive and domain factors into user-centric evaluations might prove crucial [163, 404]. This involves developing personalized interfaces tailoring explanation complexity and considering organizational and social contexts like trust and collaboration. Open repositories hosting benchmarks, frameworks, and setups can foster comparative analysis, cross-validation, reproducibility, and best practice sharing.

### 2.5.3 Future Research Directions

Based on the identified limitations, we propose a structured research agenda organized into four main directions(Table 2.13) to investigate the practical implementation and validation of XAI techniques in various PM tasks and domains.

The (1.) *foundational* direction aims to develop a unified taxonomy and conceptual framework for understanding and classifying XAI techniques in PM. The (2.) *methodological* direction focuses on adapting existing XAI techniques and developing novel methods tailored to the specific challenges of PM data and algorithms. The (3.) *evaluation-oriented* direction emphasizes establishing standardized benchmarks and user-centric evaluation methodologies considering cognitive and domain-specific factors influencing explanation effectiveness. The (4.) *application-focused* direction investigates applying and validating XAI techniques in underexplored PM tasks like conformance checking and process enhancement, as well as conducting case studies and action research in various domains like healthcare and manufacturing.

Still in terms of PM tasks, especially for onformance checking, future research should develop explainable techniques providing meaningful insights into deviations between observed behavior and reference models [21, 334]. This may involve novel alignment-based techniques and interactive visualization tools. While for process en-

**Table 2.13:** Future research directions for explainable AI in PM.

| Direction | Research Questions | Objectives |
|---|---|---|
| *Foundational* | - What are the fundamental concepts and principles of explainable AI in PM?<br>- How can the explainable AI landscape be characterized and categorized in the context of PM? | - Develop a unified taxonomy and ontology for explainable PM<br>- Establish a conceptual framework for understanding and classifying explainable AI techniques in PM |
| *Methodological* | - How can existing XAI techniques be adapted and extended to address the specific challenges of PM?<br>- What novel XAI techniques can be developed specifically for PM tasks and data structures? | - Develop scalable and generalizable XAI methods for PM<br>- Propose new XAI techniques tailored to the unique characteristics of PM data and algorithms |
| *Evaluation-oriented* | - What are the key dimensions and metrics for evaluating the quality and effectiveness of explanations in PM?<br>- How can user-centric evaluations be designed and conducted to assess the cognitive and domain-specific factors influencing explanation effectiveness? | - Establish standardized benchmarks and evaluation frameworks for explainable PM<br>- Develop user-centric evaluation methodologies considering the cognitive and domain-specific aspects of explanation effectiveness |
| *Application-focused* | - How can explainable AI techniques be applied and validated in different PM tasks, such as conformance checking and process enhancement?<br>- What are the domain-specific challenges and opportunities for applying explainable AI in various industries, such as healthcare, manufacturing, and public administration? | - Investigate the application of XAI techniques in underexplored PM tasks and domains<br>- Conduct case studies and action research to validate the effectiveness of explainable AI in real-world PM scenarios |

hancement research should explore explainable techniques supporting identification and prioritization of improvement opportunities based on performance indicators and stakeholder goals [137,339]. This could involve multi-criteria decision-making frameworks and what-if analysis tools. To finally address domain-specific applications, case studies and action research should validate the effectiveness of explainable techniques in real-world scenarios, collaborating with industry partners. As an example, for the healthcare domain research should investigate tailoring techniques to handle clinical pathway complexity while addressing privacy and ethical concerns [150].

### 2.5.4 Methodological Considerations for the SLR & Future Research Directions

This systematic review has followed rigorous methodological guidelines and best practices to provide a comprehensive analysis of XAI in PM. However, it is important to acknowledge potential limitations and opportunities for future research.

The search strategy, while extensive, may not have captured all relevant studies due to the rapid evolution of the field and the focus on specific databases and proceedings. Forward and backward reference checking was employed to mitigate this issue. The quality assessment, though based on established criteria and involving multiple reviewers, may have some inherent subjectivity. The heterogeneity of the included studies posed challenges for synthesis, but a structured approach and transparent reporting were used to address this. Most included studies were conducted in academic settings using public or synthetic datasets, potentially limiting insights into real-world industrial challenges.

| ***Process Mining Task*** *vs.* ***XAI Method*** | *Intrinsic Transparent Models* | *Model-Specific Post-hoc* | *Model-Agnostic Post-hoc* | *Visual Explanations* | *Local Explanations* |
|---|---|---|---|---|---|
| *Predictive Monitoring* | **9** [62, 189, 191, 383, 398, 432, 482, 496, 504] | **7** [53, 162, 189, 191, 206, 383, 496] | **24** [6, 7, 26, 27, 66, 80, 143, 162, 162–164, 207, 208, 381, 382, 390, 432, 448, 449, 480, 481, 499, 500] | **12** [27, 80, 100, 142, 163, 271, 300, 374, 449, 457, 491, 498] | **22** [6,7,62,80,122,162, 163, 169, 206, 275, 334, 374, 381, 383, 448, 449, 457, 480, 481, 491, 504] |
| *Process Discovery* | **4** [150, 189, 428, 499] | **1** [150] | **1** [21] | **4** [122, 150, 189, 428] | **2** [150, 428] |
| *Conformance Checking* | **2** [21, 340] | **0** | **1** [334] | **1** [334] | **3** [21, 334, 340] |
| *Diagnostics* | **3** [27, 94, 142] | **2** [27, 80] | **4** [80, 94, 334, 340] | **4** [27, 80, 271, 340] | **5** [27, 80, 271, 334, 340] |
| *Enhancement* | **2** [27, 339] | **1** [27] | **3** [27, 137, 339] | **1** [137] | **2** [27, 339] |

**Table 2.14:** Process Mining Tasks and XAI Methods with Citation Keys (Rotated)

# Chapter 3

# Perspectives from Practitioners on Strategies and Barriers

In the previous chapter, we undertook a SLR to analyze the state-of-the-art in XAI techniques applied to PM. The review revealed current research trends, influential contributions, and most importantly, limitations and gaps in existing approaches. The SLR highlighted technical challenges such as scalability and generalizability, evaluation-related issues including the lack of standardized benchmarks and user-centric evaluations, and application-oriented limitations like the limited exploration of XAI in certain PM tasks and domains.

Building upon these findings, this chapter presents a qualitative study that investigates the real-world strategies and barriers in explaining PM insights from the perspective of practitioners. The study design, including the questionnaire and focus group discussions, was inspired by the need to bridge the gap between the technical advancements identified in the SLR and the practical challenges of implementing explainable PM in organizational contexts. Indeed, the SLR provided a solid foundation for understanding the current state of XAI techniques in PM, but it also revealed the need for a more empirical approach to uncover the best practices and challenges faced by practitioners in real-world deployments. By probing the experiences and insights of professionals working with PM solutions, we aim to complement the technical findings of the SLR with a deeper understanding of the organizational and domain-specific factors that influence the adoption and effectiveness of explainable PM.

Through a combination of surveys and interviews, we explore the strategies practitioners employ to address the limitations identified in the SLR, such as developing scalable and generalizable XAI solutions, establishing evaluation criteria that consider

user needs and domain constraints, and tailoring explanations to specific PM tasks and application areas. We further discuss barriers and enablers of explainable PM adoption, such as data integration challenges, stakeholder engagement, and change management. Overall, the insights gained from this practitioner study are not only grounded in the technical foundations laid out in the SLR but also extend beyond them to capture the complex realities of implementing explainable PM in practice. By synthesizing these findings with the technical advancements discussed in the SLR, we aim to provide a more comprehensive understanding of the opportunities and challenges in developing and deploying XAI solutions in PM contexts.

The chapter is structured as follows. First, we describe the methodology of the qualitative study, including the participant selection, data collection, and analysis procedures (Section 3.1). Next, we present the key findings, organized around the themes of explanation strategies, evaluation approaches, domain-specific considerations, and organizational factors (Section 3.2). We then discuss the implications of these findings in relation to the limitations and research agenda of the SLR (Section 3.3), highlighting how the practitioner perspectives contribute to advancing the field of explainable PM.

## 3.1 Methodology

### 3.1.1 Research Question Definition

The primary objective of this study is to identify and analyze the strategies employed by practitioners to interpret and explain PM analyses in real-world scenarios. Our research question is: *What strategies do practitioners use to interpret and explain process analyses in practice?* In addition to uncovering the currently used strategies, we aim to understand clients' expectations before and after the analyses are conducted. This research focuses on exploring how practitioners manage the explanations of process analyses and identifying potential gaps in meeting clients' expectations. By investigating the crucial role of making process models intelligible through XAI, we seek to highlight the impact of these strategies on organizational effectiveness and competitiveness.

The study emphasizes two key aspects: interpretability and explainability in PM. "*Interpretability*" refers to the ability to comprehend the structure and outcomes of process models, including the capacity to examine and understand the underlying mechanisms and results of process analyses. "*Explainability*" encompasses the ability

to present PM results in a clear and accessible manner, enabling clients and stakeholders to grasp the insights derived from the data. This involves not only visualizing the process model but also interpreting its implications, clarifying the reasoning behind specific findings, and suggesting future actions based on these insights.

While adoption strategies are also considered, they serve primarily as contextual information to help understand the challenges and opportunities associated with the core topic of explainability in PM. In this study, "*adoption strategies*" refer to the approaches organizations use to integrate and leverage PM techniques in their operations.

### 3.1.2 Research Phases

To generate generalizable insights, we employ a two-phase qualitative methodology. The first phase involves the development and dissemination of an online questionnaire to collect data on the empirical strategies currently used in the field [171, 257]. The survey consists of 34 questions, each meticulously designed to address a specific aspect of our research. *Demographic* questions are included to gather information about participants' professional and organizational backgrounds. A set of questions is dedicated to understanding their methods for analyzing process models, referred to as *inspection strategies*. The majority of the questions focus on how participants communicate the results and implications of process analyses, which we term *explainability strategies*.

Drawing from relevant literature and our understanding of the field [5, 294, 313, 399, 419, 467, 469], we anticipate a wide range of responses for each question and design the possible answers accordingly. However, to capture unforeseen insights, we include "*other*" options and open-ended questions, allowing for a comprehensive understanding of participants' strategies and experiences. Figure 3.1 provides a visual overview of the questionnaire structure.

The second phase of our research involves conducting online focus group interviews, during which participants can expound on their viewpoints. These interviews are informed by preliminary findings from the questionnaire, thereby creating a feedback loop between the two research phases. By identifying prevalent practices through the questionnaire and then drawing out overarching themes via grounded theory [77] in the second phase, our methodology provides a robust understanding of the specific tactics practitioners employ.

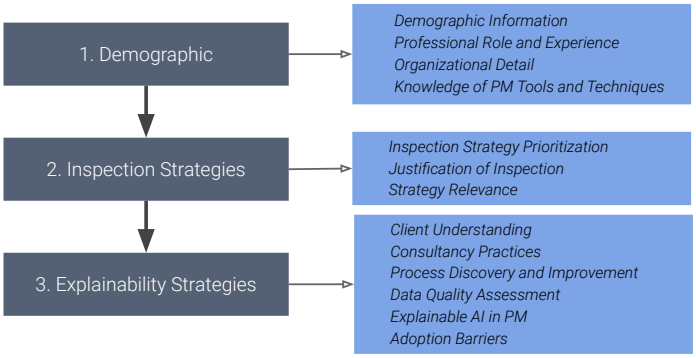


**Figure 3.1:** Partition of the questionnaire: showcasing central themes across the three sections.

### 3.1.3 Participant Selection

Our participant selection strategy was designed with an emphasis on diversity and richness of insights. We sought individuals who met specific selection criteria:

- *Volunteers:* We invited individuals willing to voluntarily participate in our research, valuing the unbiased perspectives that volunteer participation could bring;
- *Consultancy Experience:* To ensure our findings were grounded in practical, professional experience, we targeted individuals with prior or current consultancy experience in the realms of PM and BPM. This allowed us to tap into the real-world applications and challenges of deploying different PM approaches;
- *Work Affiliation:* Participants were selected from varied work affiliations – from startups to multinational corporations – across different industry sectors. This aided in capturing a broader view of PM usage across different organizational structures and cultures;
- *Research Experience in BPM or PM:* Inclusion of individuals with a background in BPM or PM research ensured an academically rigorous perspective, enabling us to bridge the gap between theory and practice.

By adopting these selection criteria and reaching this diverse audience, we carefully chose distribution channels for our questionnaire. These included professional social media platforms, forums, and mailing lists that were relevant to our study and that had the potential to reach our target respondents. We distributed our questionnaire

through multiple channels to minimize sampling bias: 60% through professional social media platforms (LinkedIn groups related to PM, BPM, and RPA), 10% through forums (like IBM Community Process Mining, Reddit r/automation), and 30% through private mailing lists. This strategic distribution helped us gather rich, diverse, and knowledgeable insights into our research question from 54 of the 640 individuals contacted, yielding an 8.44% response rate. The second research phase, consisting of the follow-up interview, was conducted with willing participants in February 2023.

### 3.1.4 Design & Coding

Our questionnaire was designed after a thorough review of related literature, clear delineation of our research scope, and thoughtful drafting of questions to draw out significant responses. Each question was crafted to correspond to a distinct facet of our study: demographics questions to document the professional and enterprise background, others to probe into inspection strategies (understood as the methods used to dissect and understand the process models) and explanation strategies (seen as the ways used to articulate the results and implications of the process analysis). The answer choices for each question were crafted to include the expected range of responses while permitting unforeseen insights through open-ended options and questions.

The second phase of our research utilized a rigorous grounded theory approach [77], which is a systematic methodology in the social sciences involving the construction of theory through the analysis of data, also consulting previous work deploying this approach within engineering or business studies [172, 345, 451]. This approach was executed using NVivo, a qualitative data analysis (QDA) software. NVivo facilitates a rich and deep exploration of data to uncover themes and patterns. It provides an organized system for data coding, sorting, and categorizing, thereby enabling an in-depth, nuanced analysis.

Grounded theory as a method involves the constant comparison of data to identify common threads and variations. Our research deployed three iterative stages of this method: open coding, where data (from the survey and interviews) were examined to identify and label significant segments; axial coding, which involved connecting categories to their subcategories, relating categories at property and dimensional levels, and identifying new categories and themes; and finally, selective coding, which integrated and refined these categories to build a coherent, holistic theoretical framework.

#### Ethical Concern & Material

Ethical requirements were strictly adhered to during the entire research process. Our questionnaire was administered using Microsoft Forms, respecting the principles of voluntary participation, confidentiality, and data security, in compliance with EU GDPR guidelines. The complete structures of our questionnaire and interviews, as well as our coding strategy and the process by which categories emerged, are detailed in the Appendix.

## 3.2 Results

Of data gathered, Sect.4.1 reports descriptive findings and outcomes from the questionnaire. Sect.4.2 elaborates on those findings by the interviews to outline the major themes emerging in the qualitative coding analysis. The overall study flow approach is reported below in Figure 2 while subthemes emerged from the interviews are presented in Table 1 at the end of this section.

### 3.2.1 Questionnaire Findings

#### Demographics and PM background

The questionnaire was designed with grouped sections, starting from the respondents' PM backgrounds. Despite our efforts to diversify the sample, the majority of them were *men*[35 and 19 *women*], at least *graduates* [16 Bachelor's degree, 29.63%; 28 Master's, 51.85%], with a *STEM background* [36; 66.67%], and primary professional roles of *business consultant* [16; 29.63%], *process engineer* [7; 12.96%] and *analyst* [6; 11.11%], *data scientists* [6; 11.11%] and *researchers* [5; 9.26%]. The organizations of respondents were tendentially above the *medium size* (150+ employees)[9; 16.66%] with a pronounced peak in *corporations* (5000+)[16; 29.63%], based in continental *Europe* [36; 66.67%] and *America* [14; 25.93%], consistently in the *Consulting & Communication* [22; 40,74%] or *Business & Finance* [12; 22-22%] domains. Two third of organizations were not PM vendors.

On average organizations approached PM *recently* (13 years)[21; 38.89%], even if respondents witnessed to have on average *equal* [14; 25.93%] or *more* (35 years [18; 33.33%], 510 [11; 20.37%], 10+ [6; 11.11%]) *personal experience*. In terms of the frequency of PM software tools deployment in job tasks respondents reported *daily* [18; 33.33%] or *weekly* use [10; 18.52%], followed by longer time periods, likely related to the clients' nature of project contracts. The most common PM software

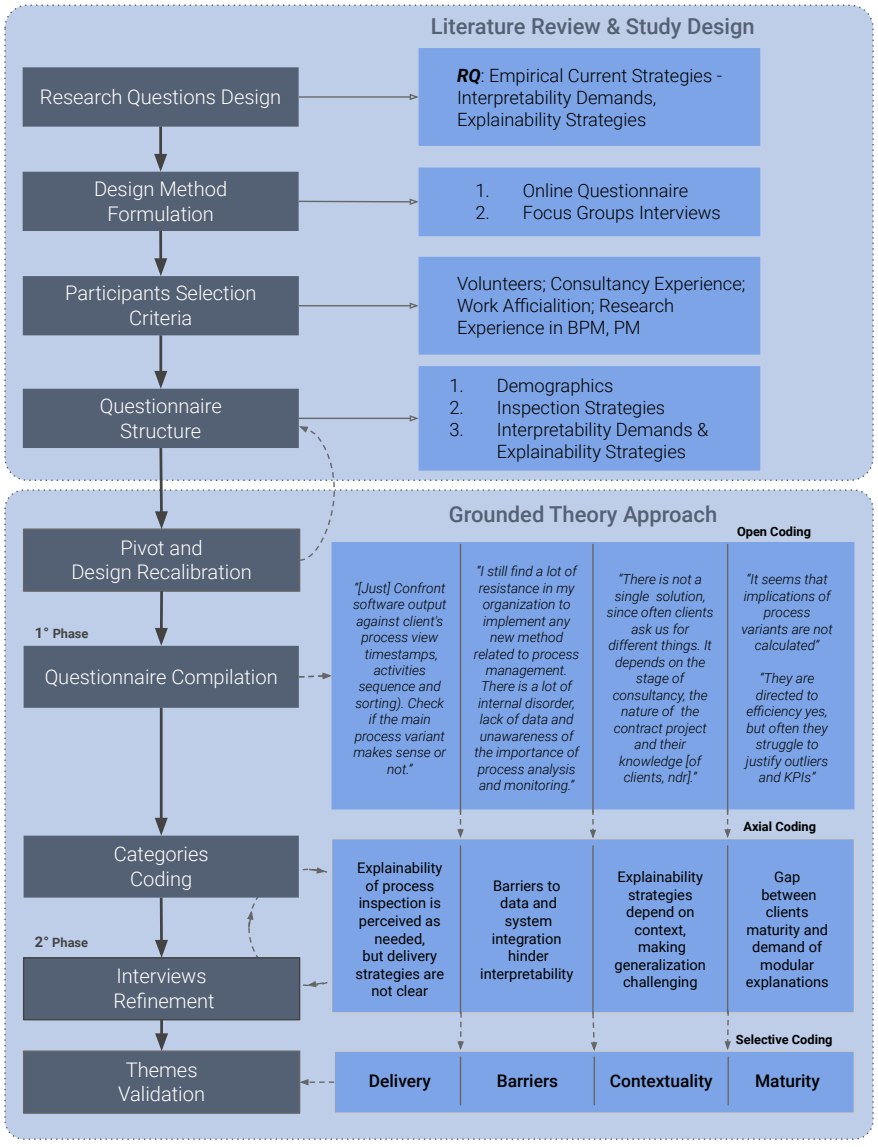


**Figure 3.2:** On the left, flow representing analysis procedure adopted. On the right, the four emerging thematic categories from GT codings

were *Celonis*, *ProM*, and *Apromore*. Other mentions included *Fluxicon*, *Minit*, *UiPath*, *SAP Signavio*, *Fortress IQ*, and *Abby* among others. Some reported using in-house developed tools or software not listed, such as *ClickUp*. Regarding knowledge of PM notations, practitioners seemed to ignore the denomination of *Directly-Follows*

*Graphs* (DFGs) [28% unknown] and *Petri Nets* [32%] despite their deployment among PM vendor solutions, while *BPMN - Business Process Modeling Notation* [63% more than advanced] and *Process Trees* [32%] were most known.

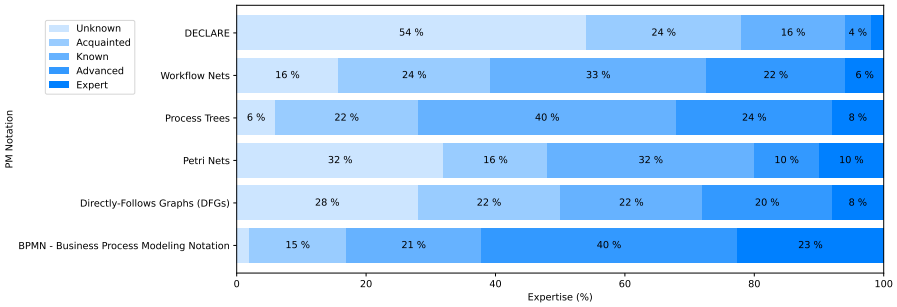


**Figure 3.3:** Knowledge of PM notations

## Inspection strategies

Practitioners highlighted that their initial focus when analyzing a process model was on *activity patterns* [24% 1st choice; 10% 2nd; 31% 3rd], followed by *KPIs* [33.3% 1st choice; 18% 2nd; 4% 3rd], and *loops* as reworks [18% 1st choice; 18% 2nd; 14% 3rd]. Notably, temporal profiles like *KPIs evolutions among activities* and *temporal separations* were ranked lower in priority, but their significance was underscored in relation to reworks and process variants for understanding process behaviors. In open-ended feedback, practitioners emphasized the importance of identifying conformance issues and optimization opportunities. Some practitioners noted that verifying conformance and existing documentation is essential, as *"it quickly provides insights to clients, building their confidence"*. KPI analysis helps understand process performance and pinpoint optimization areas. Examining activity patterns and process variants enables practitioners to *"discern primary process paths, detecting potential bottlenecks and inefficiencies"*. Identifying outliers can reveal process complexity and areas needing further analysis. Assessing loops and rework provides insights *"into possible waste"* and improvement areas. Lastly, analyzing cycle times allows practitioners to recognize opportunities for cycle time reduction, such as between critical activities or intermediate cycle times.

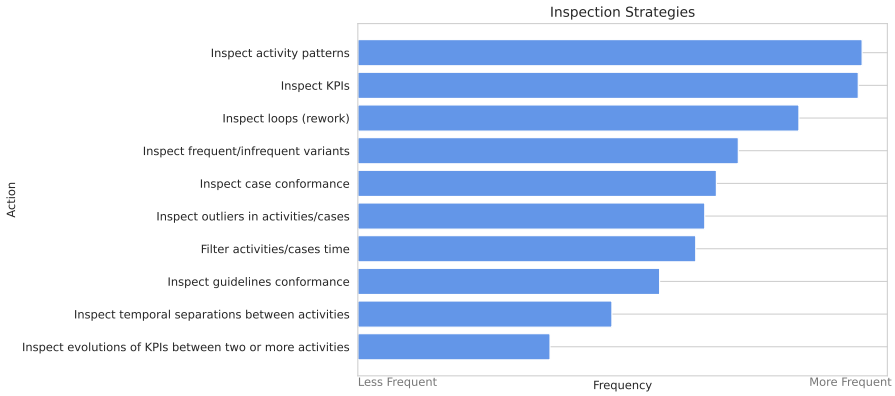


**Figure 3.4:** Priority actions taken by practitioners when inspecting a process model

### Interpretability demands from clients

After inspection, we posed questions to practitioners regarding clients' demands, highlighting their needs and expectations within PM. For the maturity of clients, it appears that the majority of clients [25; 46.30%] have been *introduced* to PM opportunities, while only a smaller percentage have *sufficient* [6; 11.11%] or *advanced* [3; 5.56%] *knowledge* of general PM notions. Given also clients that *never heard of PM before* their consultancy [14; 25,93%], results denote still a significant portion of clients unfamiliar with PM concepts.

As in line with inspection strategies, the client's type of PM project requirements [467] reflected a strong ambivalence between *performance oriented*(for KPIs optimization) [34.2% 1st; 28.9% 2nd] and *explorative* (get to know organizational processes with no previous process models) [36.8% 1st; 15.8% 2nd] while mature process projects lagged behind, such as ones *goal oriented* (root cause analysis) [15.8% 1st choice; 42.1% 2nd] and of *process monitoring* [13.2% 1st choice; 13.2% 2nd]. Coming down to specific task requirements from clients, *process simulations*, *process predictions*, and *comparative variant analysis* have lower engagement rates compared to other tasks such as *process model inspection*, *process performance measurements*, and *process enhancement*. Guidance also appears to be less frequently engaged. Yet, these tasks may still be important for process improvement and control, monitoring and evaluation, and guidance and support. Later in interviews, we inquired why these tasks have lower engagement rates and whether there is a lack of resources, understanding, or emphasis on these tasks within the organization or among the participants.

In terms of clients' expectations it was majorly reported that clients detect *some* [25; 46.30%] or *severe* [7; 12.96%] inconsistencies with their previous theoretical processes. A few practitioners reported that they often *see their true processes aligned* with their previous theoretical processes [11; 20.37%], while others indicated that clients *did not have any prior knowledge* about their processes [8; 14.81%]. The remaining respondents had a variety of other experiences, e.g. clients refusing to accept the process analysis. This suggests that there may be a gap between theoretical and actual processes in many organizations. Among the most frequently cited unexpected aspects of the process stood *number of variants* [37; 68.52%] and *loops* [28; 51.85%]. This suggests that analyzed processes may be more complex and varied than previously thought. *Reworks* [29; 53.70%] and *unexpected execution time* (e.g. throughput, output time) [27; 50%] were also frequently cited as unexpected aspects of the process. This suggests that there may be issues with efficiency and productivity in the processes being analyzed. Other unexpected aspects nominated were *internal activities* (e.g. billing, blocks) [20; 37.04%], *resource usages* [17; 31.48%], *KPIs values* [14; 25.93%], *data format and requirements* [11, 20.37%], and *stakeholder's activities* [9; 16.67%].

### Explainability strategies

On consultancy medium, the majority of the responses indicate a *hybrid approach* [23; 25.93%] i.e., direct consultancy is given at the beginning of project analysis and then monitoring activities are offered through indirect tools are used, having consultancy available on request. Closely related, *full direct* consultancy[18; 33.33%] is provided at the beginning of project analysis and then consultancy for monitoring activities occurs with less frequency. The rest of the responses indicated either that no direct consultancy is provided but is *available upon request* or that the consultancy is *only provided to internal teams* and not external clients. In terms of medium frequency indeed *direct consultancy* was mentioned by 35 respondents [61.11%]. Services such as *personal BI dashboard* and *informative reports* were also common, mentioned equally [21; 38.99%], followed by *self-assessment tools* and *automated notifications*, such as emails and text messages [13; 24.07%, equal]. *Question-answering systems* such as chatbots were the least mentioned [6; 11.11%].

When asked about medium efficiency, i.e., the most usable and less time-consuming, respondents reported in a ranking list that *direct consultancy* is the primary choice [38% 1st choice; 13% 2nd; 3rd 17%], followed by *informative reports* [8% 1st choice;

38% 2nd; 29% 3rd], and *self-assessment tools* [25% 1st choice; 17% 2nd; 19% 3rd], followed finally by *personal BI dashboard* [15% 1st; 15% 2nd; 25% 3rd; 27% 4th] and lastly by *automated notifications* [19% 4th; 19% 5th; 29% 6th] and *question-answering systems* [8% 4th; 40% 5th; 42% 6th]. This hints at the fact that fully automated text solutions with no interventions from the practitioners, e.g. NLG template-based conversational architectures, might be perceived as the least suitable form of explanation. This tendency was confirmed when asked about how information was delivered in indirect mediums outside of direct consultancy. Majority of participants claimed to *not* offer such tools [18; 33.33%], while others make the information *always available on their platform* [10; 18.52%]. Some organizations *automatically send the information* every fixed period of time [5; 9.26%], or when certain key performance indicators are achieved [3; 5.56%]. Others send the information *reactively upon business requirements*, e.g. when anomalies are detected using streaming PM techniques.

Concerning cognitive aids to elaborate visual process insights, respondents replied as a first instance to favor approaches to address the *modeling element interactivity* (e.g., filtering, and event abstraction)[25, 46.30%]. On a similar level, approaches to *change the overall presentation* [19; 35.19%] such as changing interface layout appearance or enhancing secondary notation, were closely followed by approaches to support learnability [15, 27.78%] such as worked-example effect, comprehension guidelines, tool support. The last approaches considered regarded *modularized process* [8; 14.81%] (e.g. horizontal, vertical, orthogonal) while 7 respondents [12.96%] marked all of the options. For data quality assessment to prevent technical biases, a considerable part of respondents [21; 38.99%] do not currently perform *any* data quality assessments to prevent biases. Some respondents adopt strategies such as *preventing or mitigating coverage errors* [19; 35.19%] and benchmarking construct validity [8; 14.81%]. One respondent stated that data quality issues "*tends to come out in root cause analysis*", while others specified performing manual and sample-based validation. Interestingly, concerning explainable guarantees over analysis estimates such as XAI methods in predictive process analytics, the majority of respondents *do not provide a guarantee that analysis estimates are explainable* [23; 42.59%]. Some respondents indicated that they provide explanations *upon client's request* [15;27.78%], while a smaller percentage of respondents indicated that *favor interpretable models* [7; 12.96%]. Only a small percentage [2; 3.70%] indicated to *favor black-box AI models* for their performance.

Finally, redirected to barriers hindering PM solution adoption in clients' enter-

prises: the most important barriers were found in the *lack of data quality* (i.e., no strategies to integrate and standardize data sources, missing values), being considered as “Crucial" by 25.9% and “Highly Relevant" by 31.5% of respondents. Close to that, 18.5% and 44.4% of them considered respectively as “Crucial" and “Highly Relevant" the *lack of system integration* (e.g. being inaccessible, not documented, distributed), while respectively 18.5% and 40.7% of these values were found in the option *lack of communication/understanding among internal departments*. A substantial portion, was addressed towards the *lack of data ownership* [11.1%; 38.9%], followed by gaps in the *client's expertise over process monitoring and other PM analyses* [20.5%; 34.2%]. Surprisingly, practitioners reported *not or somehow relevant* a *lack of ROI targets definition* [39.6%]. This might be due to investment strategies of clients being perceived as not directly affecting the quality of PM analyses, even if ROIs might affect organizational plans to structure data and systems (thus contributing to lower adoption barriers).

### 3.2.2 Interviews Findings

#### Explainability of process inspection is perceived as needed, but delivery strategies are unclear

Participants that agreed to expand over questionnaire findings were later contacted in February 2023. The pool of respondents widened from experienced professionals to Ph.D. researchers and professors in PM. Interviews focused on preliminary categories obtained through open GT coding by reproving, expanding, and detailing corresponding subthemes. We first detected that, in terms of delivery strategies, practitioners remarked on the usefulness of delivering insights effectively, yet communicative approaches are loosely conceived.

(1) **Objectivity analysis:** Practitioners agreed on providing accessible and objective explanations to non-technical stakeholders to ensure PM insights are effectively communicated, understood, and valued. As one noted, *“People often favor exciting technologies when, often, it is the mundane decisions and policies of an organization that provide the best opportunities for improvement.”* An evidence-based approach to process performance analysis, as well as the potential for surprising insights and excellent ROI for little effort, pair objectivity to trustworthiness, offering *“an antidote to bring evidence through process science to debunk wrong beliefs.”* Another stressed to address the root causes of issues rather than *“just the symptoms”*, starting *“from the bottom, not from the roof; tackle simple problems, even if boring, and consolidate.”*

(2) **Diplomacy and impartiality:** Practitioners should adopt a diplomatic approach when presenting results, as poor performance may have been uncovered. Clients might feel discomfort realizing that their elementary activities lacked oversight, being handled incorrectly. A practitioner suggested to *"de-personalize the analyses, make it about the situation, not the people involved"*, and *"re-orientate the findings to make them opportunities rather than problems"*, ensuring that clients accept recommendations while feeling empowered.

(3) **Reflection and empowerment:** Similarly, providing critical reflection can help clients to resonate. One practitioner mentioned starting with a *"so what?"* question to prompt the client to reflect upon the business case. Spending time understanding the customer's situation was said to be *"vital"* to deliver relevant insights, associated with monitoring tools to increase confidence, making the client aware that *"the process can be monitored from now onwards"*.

(4) **Collaboration and availability:** Establishing communities of practice on business social platforms, such as Microsoft Teams or SharePoint, facilitates collaboration between practitioners and stakeholders, providing ongoing support and guidance. As noted by one practitioner, *"sometimes I share a Slack space with my clients, so we can update even on a daily basis."*

(5) **Prototyping and simulation:** Asked about the weak use of process simulations given our questionnaire findings, some pointed to *"use simple examples to demonstrate functionalities before applying them"* to clients' use case. Employing prototypes was perceived to aid in conveying the value of solutions, with some practitioners speculating on future *"generative AI services for process simulations and prototyping"*. Similarly, *"sandboxing"* was said to *"allow stakeholders to experiment"* with potential improvements in a risk-free environment. To capture executive attention, it was recommended to emphasize financial benefit forecasts in reports, such as *"stressing on revenues, spending, and saving money"*.

### Barriers to data and system integration hinder interpretability

Effective data and system integration is paramount for delivering accurate and informed explanations to clients. However, various barriers exist in achieving seamless integration. Practitioners advanced recommendations on integration frameworks, involving stakeholders, and fostering new viewpoints.

(1) **Prioritize and integrate:** Ensuring that the framework being used can adapt to various types of data is essential. Prioritizing the relevance of certain activities

along the end-to-end process helps identify which systems need integration, as one practitioner mentioned the importance of *"figuring out which systems are being used and mapping their connections"*.

(2) **Involvement and foresight:** Involving IT and data engineers early in the process is recommended to identify potential issues and overcome technical barriers. Emphasizing the end-to-end (i.e.,*"helicopter view"*) process viewpoint in contrast to the inside-out perspective can help stakeholders see the bigger picture and understand the impact of their actions on the process as a whole. A practitioner suggested *"involving the employees' organization or worker council early on"* and being aware of different standardization requirements of clients to help overcome data and system integration barriers. Another advocated for *"starting with a Minimum Viable Product outside the live environment"* to assess interpretability before investing effort.

(3) **Preprocessing tools:** Setting up a good preprocessing facility is crucial for clean and ready data integration. To strengthen that, *"relying on external tools"* can provide good connectivity options to external data sources and allow data aggregation from different sources.

(4) **Data collection and improvement:** Collecting as many types of data as possible, even if considered irrelevant to the framework, is recommended. A practitioner emphasized the value of *"keeping a list of data improvements"* to optimize processes and enhance data quality over time. Another recommended *"obtaining anonymized data drops from respective departments"* instead of connecting directly to databases for creating a proof of concept.

(5) **Brainstorm and enrichment:** Favoring brainstorming activities, such as workshop methods, can help identify important data silos and associate them with the rough process. It was noted that *"adding customer, partner, and employee experience data to internal process performance data"* can enrich and expand the insights gained. Different forms of presentation models, such as cost and resource simulations, BPMN models, and customer journey models cater to different stakeholders.

### Explainability strategies depend on context, making generalization challenging

The effectiveness of explainability strategies is influenced by the unique context of each client's organization and project. Approaches to address contextuality challenges rely on providing narratives while making clear end-goals, prerequisites, and intentions of clients is considered advisable.

(1) **Assessing organizational culture:** Conducting assessments through organizational culture diagnostic tools can help identify gaps in expertise and resources, enabling a more tailored explainability strategy. As said before, emphasis is given to tap on the client organization's language when explaining the process, since as one practitioner stated, *"Use the client organization's language when depicting and explaining the process."*

(2) **Reference stories:** Providing reference stories of previous implementations with other clients can help illustrate the adaptability of the methodology and establish the unique characteristics and needs of the client's organization.

(3) **Target vs. method:** Differentiating between PM as a target and as a method can help align the explainability strategy with overall business objectives. It is essential to clarify the purpose of implementing PM and the desired outcomes. One practitioner questioned, *"Why are we implementing Process Mining in the first place? What do we want to achieve at the end of the day?"*.

(4) **Adapting to users' intentions:** Conceiving explanations upon clients' intentions helps tailor the strategy to the specific context and stakeholder needs. Practitioners mention the importance of being careful about the impact of their messages, as different stakeholders have varying technical backgrounds, expectations, and constraints. As one practitioner advised, *"be mindful with whom you are talking within a company."*

(5) **Adapting methods:** When the gap between the expertise level of the PM team and the client's organization is due to technical or context-informed issues, it may be necessary to change the explanation method or train resources within the client's organization. A practitioner stated to solve the gap *"by forming some resources within the client's organization or making a more shallow explanation model."*

### Gap between clients' maturity and demand for modular explanations

Several strategies can be employed to address the challenges arising from the gap between clients' maturity levels and their demand for modular explanations.

(1) **Training and introductory sessions:** Practitioners should provide training and introductory sessions on PM to familiarize clients with the technology. One practitioner noted, *"The majority of clients do not know what is PM, an introduction to it in which its potential is presented may be useful for improving their trust."*

(2) **Validation and feedback:** Involving clients in process walkthroughs, data validation, analysis, and insights validation can help to teach about and engage them with the process. As one practitioner mentioned, *"Ask middle-term feedbacks using*

*pop-up to double-check if they are into the procedure"*. For practitioners, to evaluate the effectiveness of their explanations, it was advised to collect feedback through comments, surveys, polls, and assess whether clients continue working on the proposed solutions.

(3) **Filtering information:** Presenting only meaningful and relevant data can help maximize clients' limited attention. One practitioner said to consider *"cognitive ergonomics of GUIs - guide them to avoid information overload"*, while another explicitly warned to *"use the limited attention of the client wisely"*.

(4) **Addressing varying client expertise:** Clients may have different levels of technical expertise or domain-specific knowledge. Tailoring explanations and communication strategies to suit each client's unique background is essential for effective engagement. A practitioner shared, *"be mindful with who you are talking with in a company since they have different technical backgrounds, different expectations, and constraints"*.

(5) **Management assessment:** Conducting an objective and comprehensive business process management (BPM) maturity assessment can help practitioners adapt their approach to clients' needs. A practitioner noted the importance of *"undertaking an objective and comprehensive BPM maturity assessment and combine with current and planned levels"*.

1. **Delivery:**
    1. *Objective analysis*
    2. *Diplomacy and impartiality*
    3. *Reflection and empowerment*
    4. *Collaboration and availability*
    5. *Prototyping and simulation*
2. **Barriers:**
    1. *Prioritize and integrate*
    2. *Involvement and foresight*
    3. *Preprocessing tools*
    4. *Data collection and improvement*
    5. *Brainstorm and enrichment*
3. **Contextuality:**
    1. *Assessing organizational culture*
    2. *Reference stories*
    3. *Target vs. method*
    4. *Adapting to users' intentions*
    5. *Adapting methods*
4. **Maturity:**
    1. *Training and introductory sessions*
    2. *Validation and feedback*
    3. *Filtering information*
    4. *Addressing varying client expertise*
    5. *Management assessment*

**Table 3.1:** Summary of subthemes identified in coding categories

## 3.3 Discussion

Our quantitative analysis revealed that the respondents are primarily males with STEM backgrounds, engaged in business consultancy, process engineering, and analysis roles. They are mostly affiliated with medium to large-sized organizations, predominantly in Europe and America. While the respondents demonstrated significant experience in PM or BPM, their organizations reflected a lesser maturity level in the same areas. Qualitatively, we identified the importance of customizing explainability strategies in PM based on the specific requirements and circumstances of the stakeholders involved. This involves understanding the organizational culture, language, and distinct needs of the clients. A key insight from our study was the differentiation of PM as a "target" and as a "method," providing a framework for aligning explainability strategies with overarching business objectives.

Our findings also highlight the role of stakeholder involvement in validating PM findings and documenting process models. Linguistic alignment and accessible explanations emerged as critical factors in facilitating productive client-practitioner interactions. We identified potential strategies to bridge the gap between technical and non-technical stakeholders, including effective communication, training initiatives, and sharing best practices within the community. These could take the form of PM introductory sessions, workshops, data filtering mechanisms, and assessments of BPM maturity.

The SLR in Chapter 2 identified several limitations and research directions in the field of explainable PM, including the need for scalable and generalizable XAI techniques, user-centric evaluation methods, and domain-specific applications. The findings from our practitioner study provide valuable insights into how these challenges are addressed in real-world settings and contribute to advancing the research agenda outlined in the SLR.

### 3.3.1 Implications for Identified Domains and Industries

The implications of our research extend to a wide range of fields and industries, offering fresh perspectives and insights for those involved.

- *Explanation Strategies* – Our study revealed a range of explanation strategies employed by practitioners to interpret and communicate the results of process analyses. While some of these strategies align with the XAI techniques and methods reviewed in the SLR, such as the use of visual explanations and local

explanations (e.g., [164,496]), we also identified novel approaches that emerged from the practitioner perspective. For instance, practitioners emphasized the importance of tailoring explanations to specific user roles and domain contexts, which is an aspect that has received limited attention in the current literature. These findings suggest that future research on XAI techniques for PM should focus not only on developing accurate and scalable methods but also on ensuring their adaptability to different user needs and organizational contexts. As highlighted in the SLR, there is a need for more research on interactive and user-centric explanation interfaces that can support the diverse requirements of process stakeholders (e.g., [404]).

- *Evaluation Approaches* – The practitioners in our study reported using a variety of approaches to evaluate the effectiveness and quality of their explanations, ranging from informal user feedback to more structured evaluation frameworks. However, consistent with the findings of the SLR, we observed a lack of standardized evaluation metrics and methodologies for assessing the interpretability and usefulness of explanations in real-world settings. These insights underscore the importance of developing user-centric evaluation frameworks that can capture the cognitive and domain-specific factors influencing the effectiveness of explanations, as called for in the research agenda of the SLR (e.g., [163]).

- *Domain-Specific Considerations* – Our findings highlight the importance of considering domain-specific factors when developing and deploying explainable PM solutions. Practitioners reported adapting their explanation strategies to the unique requirements and constraints of different industries, such as healthcare, manufacturing, and financial services. This aligns with the application-oriented limitations identified in the SLR, which emphasized the need for more research on domain-specific XAI techniques and evaluation methods (e.g., [150, 398]).

- *Organizational Factors* – Our study revealed several organizational factors that influence the adoption and success of explainable PM initiatives, such as data quality, stakeholder engagement, and change management. These findings resonate with the socio-technical aspects of XAI adoption discussed in the SLR, which highlighted the need for considering the broader organizational context when developing and deploying XAI systems (e.g., [339]). To address these challenges, future research should explore the development of multi-stakeholder frameworks and methodologies that can guide the integration of

XAI techniques into existing PM workflows and decision-making processes. This may involve the development of collaborative tools and platforms that facilitate the engagement of different stakeholders in the design, evaluation, and refinement of explainable PM solutions.

To summarize these considerations, the Figure 3.5 illustrates a derived multi-step process recommended to integrate XAI into PM practices. The sequence of sections reflects the progression from understanding the domain (Step 1), through enhancing PM practices (Step 2), addressing data and system integration barriers (Step 3), making PM insights more accessible (Step 4), to implementing and validating XAI techniques (Step 5). The sequencing emphasizes that a solid foundation of robust data and system integration is crucial before applying XAI techniques, thereby ensuring the effectiveness and relevance of XAI.

### 3.3.2 Governance and ethical considerations.

As shown, the successful adoption of PM solutions hinges not only on technical aspects but also on user engagement and alignment with organizational processes. Consequently, it is crucial for practitioners to adopt a sociotechnical approach to explainability, fostering collaboration among explainability experts, policymakers, and practitioners. This comprehensive approach underscores the significance of governance considerations. As emphasized by the Responsible PM agenda [469], the ethical implications of providing process explanations merit further exploration. Transparent communication should govern information intake, prioritizing client interest, agency, and autonomy, while minimizing side effects such as information overload and privacy exposure [463]. Technically, our questionnaire revealed that practices to mitigate data quality bias, such as coverage error and construct validity in conformance checking, are not extensively adopted in consultancy.

In terms of clients, a primary ethical concern in PM is the potential for cognitive bias to influence the interpretation and application of the analysis. Cognitive bias can impact individual perception and judgment through mental processes, potentially leading to inaccuracies in understanding process insights and recommendations. This concern is especially relevant when clients have vested interests in specific outcomes (e.g., *confirmation bias* or *sunk cost fallacy*), making them prone to interpret data in ways that support their agenda [373].

Despite efforts to counteract technical and cognitive biases, we believe that sociotechnical tensions may arise during consultancy. Practitioners could be required

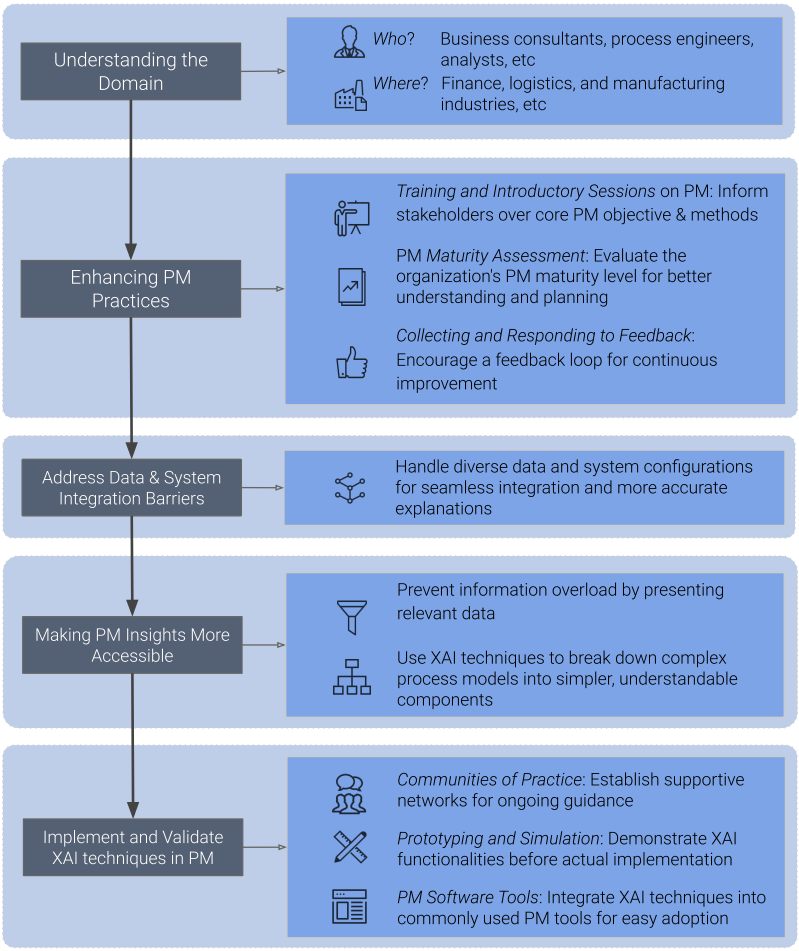


**Figure 3.5:** Strategic Steps for Enhancing PM Practices with XAI.

to provide explanations aimed at optimizing a client's business value, possibly at the expense of other ethical considerations such as data security or compliance. Ethical tensions may also manifest for practitioners, who must inform users about potential limitations in data quality, process tasks, or XAI methods utilized. This consideration is particularly critical when an explanation extends beyond technical and descriptive aspects, providing recommendations or addressing future organizational behavior within the analyzed processes [69, 233].

### 3.3.3 Limitations

While our research has inherent limitations, we have diligently worked to ensure its validity and to address potential concerns effectively. Participant recruitment and experimental setup are the main areas where limitations arise. Our sample size, though emphasizing a specific aspect of PM, represents an array of industries, roles, and responsibilities, contributing valuable insights to the field. With an 8.44% response rate, potentially influenced by nonresponse bias [424], we have successfully gathered a substantial number of responses from our target audience. The majority of participants originated from Western countries, reflecting the current prominence of PM in these regions, in terms of both service providers and demand. We acknowledge the need for wider geographical representation and will aim for this in future research. Our primary limitation concerns the focus on practitioners' perspectives, stemming from the challenges in reaching out to and interviewing PM clients as initially planned. On an individual level, this was likely due to scarce personal interest in contributing to our findings. On an organizational level, this might be also due to the fact that many companies view the implementation and use of PM as a trade secret, and therefore, they are unwilling to have their information shared publicly [176]. This might be also related to the client's fear of speaking out about their experience with PM and may not be entitled to represent their enterprises, further jeopardizing their relationship with PM providers and leading to reputational risks and competitive disadvantages [118, 212]. Thus, accessing a broader sample population of clients in the field of PM would strengthen the generalizability of the findings. Future research should account for such limitations e.g., interviewing different clients with methods ensuring confidentiality and sample heterogeneity.

## 3.4 Concluding Remarks

The findings from our practitioner study provide valuable insights into the real-world strategies and challenges of interpreting and explaining process analyses. By connecting these insights to the limitations and research agenda identified in the SLR from Chapter 2, we have demonstrated how the practitioner perspectives can contribute to advancing the field of explainable PM. Our study highlights the importance of developing XAI techniques that are not only accurate and scalable but also adaptable to different user needs and organizational contexts. We have identified research opportunities over user-centric evaluation frameworks, domain-specific XAI methods, and multi-stakeholder approaches to support the adoption of explainable PM initiatives.

While revealing, our investigation was limited to a practitioner viewpoint. To develop informed, holistic recommendations, we must understand the broader organizational ecosystem shaping responsible technology development and adoption. To lay down this ground, in the next chapter we establish a legal foundation analyzing the regulatory landscape related to algorithmic accountability and explainability. By exploring regional policy differences, we can discern how guidelines might catalyze or hinder XAI applications from our technical survey.

# CHAPTER 4

# REGULATORY STANCES FOR XAI IN PROCESS MINING

## 4.1 Introduction

In the preceding chapters, we conducted a SLR to analyze the state-of-the-art in XAI techniques applied to PM. While revealing valuable insights, this technical perspective could not fully capture practical constraints and barriers that can hinder the adoption of these techniques. To develop a comprehensive understanding, the next chapter explored limitations from an empirical perspective. Practitioners' standpoints reflected the aspiration to go beyond existing technical approaches to ensure XAI solutions for PM that are not just accurate but also usable, compliant, and ethically aligned. The questionnaire findings (Section 3.2.2) in Chapter 3 further underscore the importance of considering the regulatory sphere in the context of PM. The majority of respondents were based in Europe, reflecting the concentration of PM research and industry in this region. This aligns with the EU's position as a global leader in AI regulation, with the proposed AI Act setting forth comprehensive requirements for high-risk AI systems, including those used in PM applications. Given the stringent compliance demands and the centrality of Europe in the PM landscape, a focused examination of the EU regulatory environment is warranted.

Equipped with this conceptual foundation, we now turn to examine two complementary facets that shape the application of explainability in practice - the policy and regulatory landscape. This chapter is informed by two published papers (FaCCT'23 conference [343]; IEEE Intelligent Systems journal [342]). The first study [343] provides a comparative analysis of major policies related to AI explainability in the

European Union (EU), United States (US), and United Kingdom (UK). It identifies common themes, tensions, and limitations in current regulatory approaches across jurisdictions. Building on this international landscape, the second study [342] undertakes an in-depth thematic analysis of several articles and recitals within key EU regulations shaping legal obligations around algorithmic transparency and explainability.

Together, these interconnected studies enable a multi-faceted investigation into the nuances and complexities of the AI explainability policy landscape from both international and EU-specific perspectives. Finally, a discussion synthesizes findings from both studies, analyzing tensions between desires for transparency and barriers to full disclosure. An example scenario taken from a PM manual [399] is used to illustrate issues concretely for PM. By providing both a broad international policy overview alongside a focused EU regulatory examination, this chapter develops a comprehensive understanding of the emerging legal landscape around algorithmic accountability and explainability to impact PM adoption, informing our subsequent chapters on responsibly operationalizing explainability in PM.

## 4.2 Analysis of International Regulations and Standards on Explainability

The first study provides a comparative analysis of major policies and regulations related to AI explainability in the European Union (EU), United States (US), and United Kingdom (UK). It maps key provisions across jurisdictions and identifies common themes, tensions, and limitations in existing regulatory approaches to interpretable and accountable AI systems. By situating EU efforts within a broader global context, this macro-analysis reveals shared challenges as well as differences in strategy that characterize this complex transnational governance issue.

### 4.2.1 Methodology: a thematic and gap analysis of AI explainability policies

To conduct our research, we first source policy documents from official governmental and affiliated agencies' websites of the European Union (EU), United States (US), and United Kingdom (UK)[1]. We collect four types of documents as detailed in Table 4.1:

[1]For the European Union, we refer to official websites such as e.g., the digital strategy of the European Commission `https://digital-strategy.ec.europa.eu/`, the official online law database Eur

| Category | Description |
|---|---|
| *Communications* | Related to AI governance strategies through public statements and releases |
| *Reports* | Comprehensive studies, surveys, or official research papers that provide in-depth research information |
| *Regulations* | Legally binding rules and guidelines that dictate how organizations must behave |
| *Standards* | Technical specifications detailing implementation for AI explainability policies |

**Table 4.1:** Policy Sources

We select the documents to review based on their level of relevance and availability of the data (e.g., document content under drafting might not be disclosed to the public, but titles and expected releases might be). We consider documentation produced from 2018 onwards to ensure that policies are tackling current AI developments. For the same reason, we consider the most up-to-date versions of the documents. We exclude contexts where explainability or interpretability are presented outside of direct AI system involvement, e.g., when a regulation uses the terms in auditing procedures, thereby intending explanations as being required as justifications between humans over business conduct, etc.

We examine the regulatory landscape in the EU, US, and UK, exploring the distinct ways in which each jurisdiction has been addressing the regulation of AI over time, with a specific focus on explainability. Our thematic analysis identifies common themes across the collected documents encompassing not only "*explainability*" but also associated concepts such as "*transparency*" and "*trustworthiness*", while reporting similarities and differences in how they are accounted for. The last stage of our research is to conduct a gap analysis. While comparing themes across documents already reveals a number of gaps, we complement this understanding of the policies with research publications. Based on the themes above, we search for relevant literature stemming from various research communities (algorithmic, human-computer interaction, ethics), to identify misalignment with policies, calling for future work.

### 4.2.2 Overview of Key Regulations and Standards

In this section, we provide a concise summary of the most pertinent regulations and standards related to AI explainability, focusing on the provisions that are most relevant

Lex `https://eur-lex.europa.eu/`, and standard bodies such as CEN `http://www.cen.eu/`, CENELEC `http://www.cenelec.eu/`, and ETSI `http://www.etsi.org/`. For the UK, we refer to `https://www.gov.uk/`, and related executive public bodies such as ICO `https://ico.org.uk/`, or research national institute as The Alan Turing Institute `https://www.turing.ac.uk/`. For the US, we refer to `https://www.whitehouse.gov/`, and governamental commissions such NSCAI `https://www.nscai.gov/`, or non-regulatory agencies such as NIST `https://www.nist.gov/`.

to PM.

1. Standards: The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) are the primary organizations for AI standardization activities through their joint technical committee ISO/IEC JTC 1/SC 42 'Artificial Intelligence' [217]. ISO/IEC is working on several documents related to explainability, such as ISO/IEC AWI TS6254 (expected in 2024) [218] on explainability methods for AI systems, and ISO/IEC AWI 12792 [222] on a taxonomy of AI system transparency information elements. Additionally, ISO/IEC TR 24028:2020 [221] published in May 2020 considers explainability as a mitigation measure to AI vulnerabilities and threats. The Institute of Electrical and Electronics Engineers (IEEE) has various scattered activities related to AI explainability, including standards like IEEE P7003 [439] on algorithmic bias and output interpretability, IEEE P2894 [36] (a guide on XAI techniques, application, and evaluation expected in 2024), IEEE 7001-2021 [213] on transparency of autonomous systems, and IEEE P2976 on mandatory requirements for an AI system to be considered explainable.

2. United States (US): The US approach to AI regulation has primarily relied on self-regulation [194], with the National Institute of Standards and Technology (NIST) playing a key role in developing guidance and standards. NIST has published papers on principles and methods for explainability, such as NIST-8312 [387] and NIST-8367 [60], but these have not been consistently transposed into their AI Risk Management Framework v1.0 [357]. The proposed Algorithmic Accountability Act (2022) [3] would require companies deploying AI systems to provide explanations, engage stakeholders on explainability, and document transparency and explainability measures.

3. United Kingdom (UK): The UK's approach to AI regulation is still evolving, with a focus on providing guidance and promoting industry adoption of explainable AI. The Information Commissioner's Office (ICO) and the Alan Turing Institute (ATI) have released guidance on explaining AI decisions, including a comprehensive report 'Explaining Decisions with AI' in 2020 [216]. The UK has established the Standard Hub [76], a cross-government standard for algorithmic transparency, and plans to integrate future standards into its AI innovation strategies [154, 155]. Despite leaving the EU, the UK remains

a member of European Standards Organizations (ESOs) like CEN/CENELEC and actively engages in drafting international AI standards [154].

4. European Union (EU): The EU adopts a risk-based approach to AI regulation, with the level of regulation proportional to the potential adverse consequences of an AI system [87]. The proposed AI Act draft (2021) [358] includes provisions for interpretability and transparency requirements for high-risk AI systems, although revisions [353,354,360] have weakened some explainability concepts. The European Commission (EC) plays a crucial role in shaping the regulatory framework, working with European Standards Organizations (ESOs) like CEN, CENELEC, and ETSI [1,74,89,125,136,335] to develop standards supporting the AI Act.

The analysis spotlights the EU AI Act and the proposed US Algorithmic Accountability Act, which emphasize the need for clear and comprehensive information about the purpose, design, and decision-making processes of AI systems. This includes documenting the input data, training methods, performance metrics, and ultimately, the logic behind the systems' outputs. Conversely, the GDPR zeroes in on the explainability of automated decisions with legal or significant implications for individuals, mandating the provision of meaningful information about the logic involved and the potential consequences. International standards, such as ISO/IEC TR 24028:2020 and IEEE P7001, take a holistic view, considering explainability as a means to foster trust, accountability, and transparency in AI systems across various stakeholder groups, including end-users, regulators, and the general public. These standards underscore the importance of collaboration and communication among stakeholders throughout the AI development lifecycle, from design to deployment and maintenance. The analysis further delineates the roles and responsibilities assigned to key players, such as system providers, data controllers, and users. While the EU AI Act and US Algorithmic Accountability Act place the onus primarily on system providers, the GDPR focuses on the obligations of data controllers. International standards advocate a shared responsibility among multiple stakeholders, recognizing the need for a concerted effort to ensure explainability.

## 4.3 Explainability in the EU Regulations

Building upon this high-level comparative policy analysis, a subsequent study has been conducted to analyze into the EU regulatory context through a comprehensive

review of key provisions across regulations related to explainability. Through a comprehensive review of over 50 articles and recitals within major EU regulations, it provides a granular categorization of legal provisions pertaining to explainability requirements for AI systems. The study comprehends emerging European guidelines, acts, and directives to extract themes around transparency demands and tensions. By rigorously cataloging EU policies related to algorithmic accountability, it enables detailed analysis of regulatory complexities, gaps, and directions. Together with the analysis review detailed earlier in this chapter, these two interconnected studies provide a multi-faceted investigation into the the AI explainability policy landscape.

### 4.3.1 Methodology

To bridge the existing gap in literature we undertake a policy content analysis using the *EurLex* official repository, the EU database where draft laws and enacted legislation are reported. Our primary focus is on binding legislation related to AI, data, and digital platforms. This comprehensive review forms the basis for the subsequent mapping and analysis of explainability requirements within the EU regulatory framework

Our analysis not only considers explainability requirements for AI systems, but also the legal entities responsible for their provision. We follow a thematic order, grouping EU norms around data, AI, and platforms. Starting with data-related regulations due to their chronological appearance and the debate they sparked, we then move into AI, and finally provide an overview of platform services. This structured approach allows us to thoroughly examine the regulatory landscape and its implications for explainability in AI systems.

### 4.3.2 Data Regulations

#### General Data Protection Regulation

A discussion on the explainability dimension of algorithmic decision-making (ADM) systems and their legal obligations was triggered already before the *General Data Protection Regulation* (GDPR) enactment in 2016 [174]. The debate concerned the legitimacy of a "*right to an explanation*" based on the interpretation of recital 71 and Art.22(1) on automated individual decision-making including profiling. Alongside that, Art.13(2)(f), Art.14(2)(g), and Art.15(1)(h) regard the information to deliver to a data subject whenever personal data are collected, not have been obtained or to confirm data processing procedures. The articles mention in their commas the right from the

data controller on "*meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject"* in the automated decision. A direction mention in Art.22, where are modalities to appeal to this decision and contest it are defined, empowers the data subject to not be exposed "*to a decision based solely on automated processing, including profiling, which produces legal effects*".

Researchers have explored ambiguities around the requirement to provide information on "*the logic involved*" in ADM system outcomes that profile and affect individuals. The discussion has centered on whether this requirement pertains to data processing rationale or fuller transparency including design choices. As expounded by legal scholars [115, 184], the term "*right*" to an explanation finds a single mention, within a recital, thus holding no binding power. What makes it problematic in its implementation, however, is also the casuistry to which it refers. The context in which an explanation may be required is based on cases where an ADM system operates decisions with legal effects (or the like) on an end-user without *any* human intervention ("*solely*"). This delineation likely invalidates every possible case in which a human operator contributed to the decision about the user, whether AI systems were used as decision-making support services. Regarding the limitations affecting the efficacy of Art.22, it is relevant the principle of information proportionality adopted by the EU. In Art.15(4) there is a mention of how an adversary context (i.e., litigation) jeopardizes this right to obtain information even affecting "*the rights of freedom of others*". Furthermore, norms on business secrecy[2] offer a disincentive to disclose information potentially connected to the interests of the controller. In other words, information obtained by an end-user presumably regards only personal data, excluding ADM systems rationale and the related business design choices conceived. Also, note the lack of legislative precedents for the GDPR as ruled by the Court of Justice of the EU (CJEU). Constitutional courts can establish if an algorithmic system determined a violation on a specific individual case, yet this will not constitute a general binding standard [115].

[2]References to Art. 16 of the EU Charter of Fundamental Rights; the Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use, and disclosure: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX3A32016L0943`; and the European Parliament Resolution of 20 October 2020 on IP rights for the development of artificial intelligence technologies (2020/2015(ini): `https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html`

#### Data Act & Data Governance Act

While the GDPR lays harmonized rules regarding foremost the elaboration of personal data, the Data Governance Act (DGA)[3] enables the provision of common inter-operable data spaces in strategic ICT sectors as designed by the EU Digital Compass. Close to its final draft approval after being proposed in 2020, the DGA was paired in 2022 with a further bill advanced by the EC defined as Data Act (DA)[4]. The latter outlines data transfer processes and addresses potential information abuses that could arise from contractual imbalances, potentially obstructing fair access to shared databases. While the DGA and the DA primarily ensure legislative norms for data-sharing, their direct relevance to ADM systems is less pronounced, and they could be considered peripheral to our analysis. Nevertheless, this normative framework provides insights into what constitutes transparency of public data, once again emphasizing the principle of informational proportionality as a cornerstone, as seen in the DGA's Art.5 on data reuse conditions.

### 4.3.3 Artificial Intelligence Regulations

#### AI Act Draft

The major regulatory instance on AI was proposed with the initial AI Act draft (AIA) in April 2021. This subsequently went through extensive debate and revisions, with the final compromise text was officially approved in December 2023 following trilogue negotiations between EU institutions. The approved Act is informed by the principle of proportionality of regulatory intervention for its horizontal approach to risk i.e., being applied broadly across industrial sectors rather than being domain specific. Aside from prohibited ones, this categorization spans three levels of risk - low, medium, and high - with different application limits, transparency requirements, and oversight mechanisms.

AI stakeholders are introduced alongside definitions of risk assessment (Art.9), data quality requirements (Art.10), and extensive technical documentation for authorities (Art. 11, Annex IV); further, for high-risk systems key provisions requiring measures facilitating user interpretability and human oversight (Art. 14). For interpretability, Art.13 requires instructions explaining system capabilities, limitations

[3] http://data.europa.eu/eli/reg/2022/868/oj

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2022:68:FIN

and accuracy for high-risk systems, while requiring to attach instructions for use and information to users.

Since the first draft in 2021, scholars inquired what entails the development of "*sufficiently*" transparent systems to allow end-user interpretation of the outputs and enable their appropriate use [115, 184, 444]. To mark, the criterion of "*appropriate*" (Art.13(1),(2)) with regard to the type of transparency hinted an understanding directed for legal sufficiency rather than for a generic end-user. The major issue stood in balancing model complexity, end-user expertise, and business and legal constraints. Indeed, Art.13, does not provide either standards or procedures for evaluating such balances [444], allegedly leaving the transparency assessment at the discretion of the provider of the high-risk system [115].

As for the GDPR, objections to fully interpretable design criteria find appeal in the EU directive establishing trade secret integrity and IP rights. While transparency is desirable to respect EU user rights, it is problematic to unilaterally maximize an AI system transparency when the same users are not controlled over possible misuses [28, 59]. In this regard, end-user liability can be ascertained from the need, expressed in Art.13 and recital(47), for instructions from the provider to the end-user on the intended use. In addition, Art.13(1) disregarded the possibility of using interpretability as a burden of proof by third parties affected by the AI system.

After the final draft delivered by the Permanent Representative Committee in November 2022, the European Parliament voted in mid-June 2023 to approve its negotiating position, including notable amendments for guaranteeing algorithmic transparency and individual explainability rights. Even if some amendments of June 2023 were not later retained[5] the final version of the Act, signed into law on June 13th 2024 [131], mandates certain transparency provisions, notably targeting "*high-risk*" systems as legally required to ensure embedded explainability capacities by design under Title III, Chapter 2 [130]. We comprehensively examine key articles as follows.

**Article 13** – requires understandable AI outputs, capabilities, and instructions enabling oversight. 13(1) mandates sufficient transparency to interpret outputs for ap-

[5]The draft originally proposed in April 2021 by the [126] did not comprehend any explicit article regarding a right to an explanation (RTE). The first inclusion of this concept was proposed as *Article 69(c)* through the amendments by the [134] (JURI) in their Opinion of September 2022. These amendments were not accepted in the common position of November that year by the [93] (COREPER). It was only in the negotiating position advanced at the end of 2023 that the RTE was adopted as *Article 68(c)* by the [127]. When the EU AI Act agreement reached in December 2023 between the [128] and the Council was endorsed by members state in February 2024 [129], Article 68(c) was retained. Although, certain provisions were modified - e.g., Article 13 to direct interpretability requirements towards AI *providers* and not *users* anymore. Subsequently, in the Corrigendum released in April 2024 [130], the numbering of the "*right to explanation*" article was changed from 68(c) to *Article 86* and Article 52 to 50.

propriate use. This explains recommendations or decisions per Article 86(1). Clause 13(2) requires clear, complete documentation aiding comprehension of characteristics. As per clause 13(3)(b)(ii), documentation must elucidate accuracy metrics and impacts - inaccurate/biased systems need greater elucidation. For relevant systems, clause 13(3)(b)(iiia) also needs transparency on explanatory abilities, while 13(3)(d) requires explanations on oversight measures under Article 14 for deployment monitoring.

**Article 14** – also targets effective human oversight over AI systems via interpretability, accountability and control. Clause 14(1) requires high-risk AI systems allow for such oversight facilitating user autonomy. More specifically, clause 14(3)(c) necessitates transparency enabling accurate interpretations of outputs produced while clause 14(4) stipulates understanding capacities, constraints etc. Recital 48 supports this navigate appropriate, contextually sound AI adoption preventing risks.

**Article 50** – stresses disclosures for certain AI systems directly interacting with individuals. Clause 50(1) requires clear notifications when conversing with an AI system for sound transparency norms that respect personal dignity. For biometric categorization or emotion recognition applications, clause 50(3) asks for transparent processing explanations respecting privacy rights. Beyond design phases, Article 26(5) places transparency burdens upon AI system deployers too for monitoring best practices adherence including around accuracy metrics communicated (which clause 13(3)(b)(ii) requires providers detail initially).

**Article 86** – indeed named "*Right to explanation of individual decision-making*" allows individuals to contest opaque determinations by demanding understandable outcome explanations if delegated decisions lack transparency, as recommended by recitals 107 & 171. However, as detailed by Figure 4.1 it is subject to certain limitations:

- It only applies to decisions made using high-risk AI systems listed in Annex III, with several carve-outs under Article 6(3) for systems that do not pose significant risks to health, safety, or fundamental rights.
- The RTE exists only if a decision produces legal effects or similarly significantly affects a person in a way that they consider to adversely impact their health, safety, and fundamental rights.
- Article 86(3) specifies that the RTE applies only to the extent that it is not otherwise provided for under Union law, such as the GDPR's RTE arising from

Articles 15(1)(h) and 22(1).

In addition, Chapter V of the EU AI Act introduces specific transparency obligations for providers of general purpose AI models. Article 53(1) requires providers of general purpose AI models to maintain technical documentation (Annex XI), provide information and documentation to downstream providers integrating the model (Annex XII), establish policies to ensure compliance with EU copyright law, and provide a summary of the content used for training the model. In this line, Article 55 imposes additional obligations on providers of general purpose AI models with systemic risks, such as conducting model evaluations, assessing and mitigating systemic risks, reporting serious incidents, and ensuring cybersecurity protection.

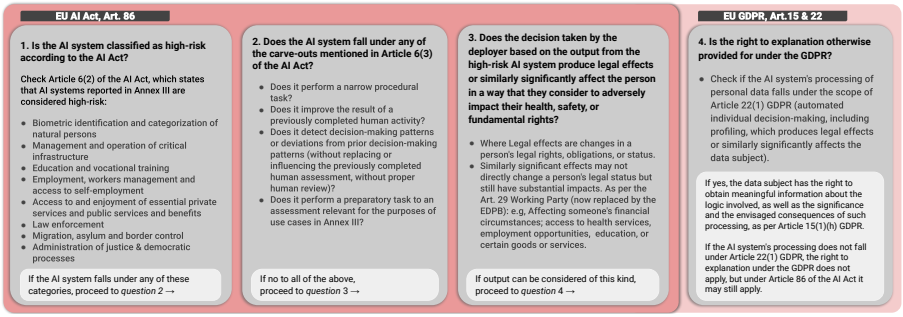


**Figure 4.1:** Prerequisites to ensure applicability of Article 86 under the EU AI Act for high-risk systems

### AI Liability Directive

In September 2022, the proposed directive on the adoption of a system of civil liability for AI (AILD) [359] was issued to better define those areas of legislative uncertainty during litigation for AI.

The directive defines mechanisms for defining provider liability with respect to a generic user, once established the impossibility for them to obtain an explanation or other justifications following the damages produced by the effects of an AI system (not only high risk). In this scenario, the user is a plaintiff, referred to as a *claimant*. Art.3 establishes that once sufficient evidence about the opaqueness of a system has been alleged - including the complexity or automation of the architecture - the claimant turns to their national court to appeal for an audit process.

This core provision introduces a rebuttable presumption of causation if three conditions are met: 1) the defendant failed to meet a relevant duty of care such as AIA

obligations; 2) their fault likely influenced the AI output or failure to output; and 3) the output or lack of output caused the damage. This presumption facilitates establishing liability in AI-related cases by connecting the defendant's actions to the AI system's role in the damage. However, it is limited to fault-based scenarios. The Directive does not address yet strict liability when AI systems function as designed but still cause harm[6]. The court is empowered to access AI system documentation and design, as stated in Art.4, to validate what is referred to at the time of the inspection as a *presumption of causality* between the damage produced and the system design. Interestingly, Art.4(2)(b),(c) refer to Art.13 of AIA if providers do not meet transparency and oversight requirements for AI system design and development. Nevertheless, procedures for assessing the sufficiency of requirements are not made clear. Recitals 27-29 specify how explanations relate to proving causation for opaque AI systems. The duty of care of an AI provider focuses on demonstrating to a mandated supervisory body that the system was compliant only with its instructions for use and documentation. This likely invalidates the possibility for end-users of interpreting system outputs or receiving explanations as a burden of proof under litigation [55].

### 4.3.4 Platform Services Regulations

#### Digital Service Act

The DSA[7] defines intermediary due diligence obligations and conditions for liability exemptions related to digital online services, including platforms for online shopping, content-sharing, cloud and messaging services, and marketplaces. The DSA distinguishes between intermediary services, hosting services like online platforms, and "Very Large Online Platforms" (VLOPs).

To enhance transparency in content moderation, the DSA requires all intermediary services to designate a legal representative and describe their methods, including use of algorithmic decision-making systems. Providers must also offer notice mechanisms for allegedly illegal information (Art.9 & 15) and clearly explain terms and condi-

[6]The proposed revisions to the Product Liability Directive (PLD) aim to complement the AILD by updating the rules on strict liability for defective products in light of emerging technologies like AI [133]. The PLD explicitly includes AI systems within its scope, widens the concept of "damage" to include data loss or corruption, and considers factors like self-learning and cybersecurity vulnerabilities when assessing defectiveness. AI providers can be held strictly liable for defective AI products and cannot rely on certain defenses if the defectiveness relates to aspects under their control, like software updates. The PLD's implementation timeline aligns with the AI Act, so businesses should consider how the two directives interact, keeping in mind the PLD will need to be transposed into national laws.

[7]https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065

tions for managing third-party content (Art.14). For adversiting, recital 68 enhance transparency for users in platforms service through "*meaningful explanations*" around the ad and such profiling referring to the GDPR. Art.26 requires online platforms to provide transparency into how users are advertised, VLOPs using recommender systems (Art.27) are subject to audits on activities like profiling and targeting recipients, having to explain "*design, the logic, the functioning and the testing of their algorithmic systems"* (Art.40(3)). Similar to that, recital 141 calls for the Commission to be able to request documentation and explanations about algorithmic systems from all providers. Art.69(2)(d) and (5) grants the Commission authority during inspections to examine algorithms and require platforms to explain system functionality, data practices, and business conducts. Additionally, Art.72(1) enables the Commission to monitor compliance of VLOPs and search engines by ordering access to databases and algorithms, and requiring explanations relating to them also through appointed experts (Art.72(2)).

While these articles focus on oversight and compliance rather than mandating interpretability, they provide authorities tools to request information and explanations about VLOPs' algorithmic systems and data practices, enabling investigation of AI systems even if it does not prescribe specific explainability standards.

#### Digital Markets Act

While the DSA is focused on regulating online platforms, the Digital Markes Act(DMA)[8] aims to regulate the access of companies to digital markets and services. Specifically, it seeks to prevent companies from abusing their position in the market by imposing unfair conditions on other companies in terms of gatekeeper access. In addition, the DMA puts requirements on companies to share data with competitors, which could lead to increased competition. In regards to explainability, Art.19(1),(2) and Art.21 envision procedures to conduct audits on service providers respectively for technical and business transparency.

### 4.3.5 Mapping

To present a comprehensive taxonomy of the different functions corresponding to an explanation, we report specific articles and related recitals in Table 1[9]. The ta-

[8] `https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:842:FIN`

[9] [GDPR] = General Data Protection Regulation (2016/679); [AIA] = Artificial Intelligence Act Draft (2021/0106); [AILD] = AI Liability Directive (2022/0303) ; [DSA] = Digital Service Act (2020/0361);

ble categorizes regulations by the explainability dimensions and stakeholders they address, providing a useful overview of existing regulatory requirements related to interpretability and explainability of AI systems arising from the policy content analysis.

| | | **Explainability Dimension(s)** | | | **Stakeholder(s)** | | |
|---|---|---|---|---|---|---|---|
| | Recital | Data | Process | Business | Auditor | Provider | User |
| **GDPR** | | | | | | | |
| Art.13(2)(f),14(2)(g),15(1)(h) | [60] | [~] | [~] | [x] | - | [G] | [R] |
| Art.22(3) | [71] | [~] | [x] | - | - | [G] | [R] |
| **AIA** | | | | | | | |
| Art.13(1),(3) | [14a, 47] | [x] | [x] | - | - | [I] | - |
| Art.14(4)(c) | [48] | [x] | [~] | - | - | [I] | - |
| Art.50(1) | [48, 70a] | - | [~] | [x] | - | [G] | [R] |
| Art.86 | [84b] | [x] | [x] | [x] | - | [G] | [R] |
| **AILD** | | | | | | | |
| Art.3(1) | [16-21] | - | [~] | [x] | [R] | [G] | - |
| Art.4(4),(5) | [22-30] | - | [x] | [~] | - | [G] | [R] |
| **DSA** | | | | | | | |
| Art.27(1),(2) | [68] | [~] | - | [x] | - | [G] | [R] |
| Art.40(3) | [141] | - | [x] | [x] | [R] | [G] | - |
| Art.69(2)(d),(5) | [146] | [~] | [x] | [x] | [R] | [G] | - |
| Art.72(1) | [93, 141] | [x] | [x] | [~] | [R] | [G] | - |
| **DMA** | | | | | | | |
| Art.21(1),(2) | [81] | [x] | [x] | [~] | [R] | [G] | - |
| Art.23(2)(d),(4) | [83] | [~] | [~] | [x] | [R] | [G] | - |

**Table 4.2:** Mapping of articles referring explicitly to interpretability and explainability in the EU regulations for AI and ADM systems

The table is divided into two main sections: 'Explainability Dimensions' and 'Stakeholders'. The former refers to the different aspects of AI and ADM systems that require explanation, while the latter refers to the entities involved in the AI lifecycle who either provide or receive them. For dimensions, we consult previous mapping work on the operationalization of ethical principles in the AI lifecycle by Georgieva et al. [167]. Their classification is divided into explainability targets such as *technical* for traceability and system description; *process* for decision-making process criteria; *business* for organizational decision-making processes and system design criteria. We

[DMA] = Digital Markets Act (2020/0374). For AIA, only high-risk AI systems are considered. Relevant recitals related to articles provision are reported. Explainability Dimension(s) covered in the article provision are signaled with an [x] if explicitly mentioned, while with a [~] to design potential desiderata. Stakeholders are marked with [G] = Giver of an explanation, [R] = Recipient of an explanation, and [I] = Interpreter.

build up on these distinctions to account for a contextual understanding during the due diligence phase or litigation, associated to the stakeholders involved in the AI lifecycle. Explainability requirements are also considered in the sociotechnical context of data and platforms use where an AI system can be implemented. We intend under *Data* both the temporal side of accessibility and fruition in the ex-ante coordinates (i.e., access to databases) and after processing of the AI system (i.e., data output) but also their topology (i.e., user profiling or reference sampled group). Under *Process* we refer to the architecture and capabilities of an AI system. Under *Business* we refer to business choices that designed the AI system, the acknowledgment to user of interacting with it, and also the social effects of an AI system of shaping an organization's business policy and affecting third parties. As previously illustrated in Section 3, articles and recitals may ambiguously define interpretability or explainability targets. For example, the concept "*logic involved*" in Art.22 of GDPR varies case by case, allegedly including only the processing of personal data to system design and related business choices. As a precaution, we use the *tilde* symbol to designate legislative ambiguity on *what* (target) can be explained.

Additional to the table, we define *how* (type) something can be explained. We draw on the typology advanced by Cabitza et al. [69] for their degree of sufficiency in covering the major casuistry for AI explanations. Their work differentiates explainability types as *computational* (i.e., how the algorithm produced any output); *mechanistic* (i.e., why so); *justificatory* (i.e., why the output is deemed as right); *causal* (i.e., which factors and how caused the output); *informative* (i.e., the implications of the output); and *cautionary* (i.e., the uncertainty behind an output). In line with these considerations, we view types of explainability desiderata as neither mutually exclusive nor rigidly defined.

However, we do identify certain interpretative tendencies. Mechanistic and computational types potentially encompass the enhancement of technical interpretability via XAI methodologies to describe an AI system's data output and process. Justificatory and causal explanations are desirable when a business explanation needs to be provided and supported by technical explanations, establishing a heuristic ground-truth of the system being explained. These explanations could also serve as evidence subject to cross-examination under litigation. Informative and cautionary explanations, on the other hand, add layers of semantic granularity, i.e., they advance epistemological knowledge about the system analyzed. Their teleological nature addresses future uncertainty and is more closely related to business explainability, informing users or organizations about remedies to achieve compliance or improve user services.

For stakeholders we advance three macro categories emerging from the analysis of regulations. By *User* we mean service clients, data subjects, or claimants; by *Provider* we refer to both AI providers and general deployers of ADM systems, also in platform services; with *Auditor*, we intend oversight bodies including legal persona delegated to conduct audits.

## 4.4 Use Case

Building upon the mapping of explainability dimensions and provisions in Sections 4.3 and 4.3.5 this discussion explores competing tensions around explainability situated across various contexts of use, from oversight procedures to user services and litigation dispute as shown in Figure 4.2. Balancing transparency for accountability with confidentiality concerns emerges as a central challenge. Regulators seek explainability to audit systems, while providers shield proprietary details. Users impacted by AI decisions desire recourse through understanding, but legal and technical ambiguities persist around useful explainability. We propose a functional view of explanation types [69, 167] based on these scenarios.

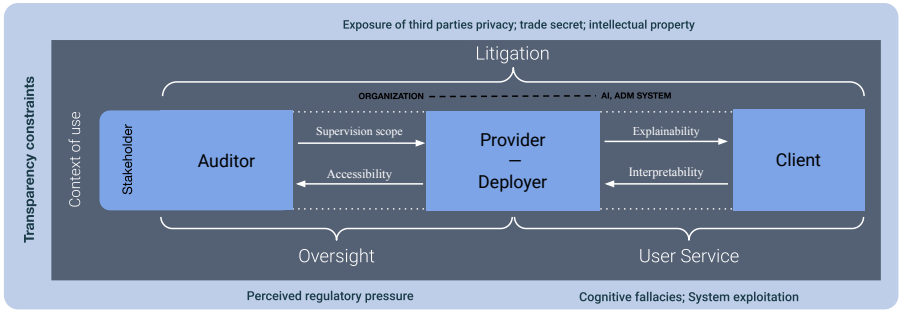


**Figure 4.2:** Contexts of use highlighted by discrepancies within different stakeholders involved in the generation and reception of explanations. On an outside layer, transparency constraints are found related to different contexts of use. An additional axis is proposed to locate the explainability area, respectively from AI or ADM systems to the business organization.

A PM case study held at Uber from [410] (Chapter 10 from [399]) illustrates how to optimize customer experience and business performance in platform-based services. By analyzing data across over 700 cities in 65 countries, Uber gained insights into process variations and identified opportunities for harmonization, efficiency gains, and customer satisfaction improvements. Under the DSA, online platforms like Uber would be required to provide clear, accessible information to users about the main

parameters used in algorithmic decision-making systems that influence the visibility, ranking, or availability of services (Article 27). This could encompass a wide range of PM insights that shape platform outcomes, such as:

1. Algorithms that match riders with drivers based on factors like location, availability, and past performance metrics derived from PM.

2. Dynamic pricing models that adjust fares based on real-time demand patterns and operational efficiency insights gleaned from PM.

3. Recommender systems that suggest optimal routes or service options to users, informed by PM analyses of historical trip data.

Providing meaningful transparency about these complex, often proprietary algorithms poses significant challenges. Overly simplified explanations may fail to capture the nuances and contextual factors shaping platform decisions, while overly detailed disclosures could compromise trade secrets or enable gaming of the system. Striking the right balance will require careful collaboration between Uber's technical teams, legal experts, and user experience designers to craft explanations that are both accurate and intelligible to diverse audiences.

Moreover, the DSA grants regulators extensive powers to scrutinize platforms' algorithmic systems (Article 40) and demand access to data and explanations about their functioning (Articles 69 and 72). This could potentially subject Uber's PM models, input datasets, and output insights to regulatory auditing and disclosure requirements. Ensuring these materials are properly documented, secured, and interpretable for non-technical authorities will be critical to facilitating effective oversight while mitigating risks of sensitive data exposure or misinterpretation.

The AI Act's requirements for high-risk systems, such as detailed technical documentation (Article 11) and user information provisions (Article 13), could further compel Uber to disclose the logic, significance, and expected consequences of its PM algorithms if their AI systems, integrated the app platform, would substantially influence contractual rights or access to services. This underscores the need for proactive, context-specific explainability measures that not only comply with legal mandates but genuinely empower affected stakeholders to understand and challenge algorithmic outcomes.

However, Uber may justifiably resist full disclosure of proprietary algorithms and sensitive data, citing trade secret protections (AI Liability Directive Article 3(4)).

Suppliers adversely impacted by PM-driven changes could demand explanations under the AI Act's user information provisions (Article 13), but the lack of precise standards for what constitutes "meaningful information" may allow superficial compliance rather than substantive clarity. The GDPR's explainability requirements for automated decision-making (Article 22) could also apply if the PM system's outputs significantly affect individual suppliers. However, the GDPR's "meaningful information" standard suffers from similar ambiguities, and its trade secret exemptions (Article 15(4)) could further limit transparency.

Developing such measures will require close collaboration between Uber's PM teams, legal experts, and impacted communities to align XAI techniques with the unique dynamics of ride-hailing platforms. For instance, explanations may need to account for real-time fluctuations in service availability, multi-sided market incentives, and network effects that shape platform interactions. Engaging drivers, riders, regulators, and municipalities in the participatory design and iterative refinement of explanation interfaces could help ensure their practical utility, perceived legitimacy, and cultural appropriateness across Uber's diverse global markets. Establishing clear protocols for users to request information about algorithmic decisions, appeal unfavorable outcomes, and seek human review will also be essential to operationalizing the DSA's user rights provisions (Article 17).

# CHAPTER 5

# ETHICAL STANCES FOR XAI IN PROCESS MINING

The prior chapter analyzed emerging regulations pertaining to algorithmic transparency and explainability, while also discussing them within regarding PM applications. As specifically seen with the Uber case, policies' emphasis on systemic transparency and regulatory oversight could significantly reshape the governance of PM in platform contexts. PM practitioners will need to work closely with legal experts, user experience designers, and impacted communities to develop context-aware explainability frameworks that balance technical sophistication with accessible communication. By doing, the previous Chapter 4 revealed tensions between moral desires for accountability and barriers to full disclosure. With a broader perspective, this facet departs from legal compliance and leads us to alignment with ethical values. Indeed, as PM increasingly leverages complex AI algorithms to derive insights from vast amounts of event log data, concerns about the transparency, accountability, and fairness of these systems have come to the fore [399, 400]. While XAI constitutes a solution to address these concerns by providing human-understandable explanations of AI-driven decisions and predictions, yet the application of those techniques, especially for the PM context, raises a host of ethical considerations and risks that shall be carefully navigated to ensure responsible deployment [313, 469].

This chapter builds on the prior analysis of state-of-the-art XAI methods (Chapter 2), empirical studies of PM explainability needs and constraints (Chapter 3), and the evolving regulatory landscape surrounding AI transparency (Chapter 4). While technical advancements in XAI algorithms and interface design are crucial, realizing the full potential of explainable PM systems requires a deep understanding of the

ethical implications and potential risks involved. Failing to proactively identify and address these ethical dimensions can lead to unintended consequences, such as erosion of trust and even harm to individuals and organizations [120, 123, 330]. The goal of this chapter is then to provide a comprehensive examination of the key ethical considerations and risks that need to be accounted for when applying XAI techniques in the PM context specifically. By synthesizing insights from ethical theories, applied ethics frameworks, and PM industry case studies, we aim to surface the context-specific challenges and tensions that can arise when striving for process transparency. This analysis will lay the groundwork for a proposed socio-technical framework for responsible PM explainability in the subsequent chapter.

The importance of this context-specific ethical analysis can not be overstated as PM deployments continue to scale across industries and geographies [400]. It is not secret that, from healthcare processes to financial auditing to manufacturing operations, PM systems are increasingly being used to inform high-stakes decisions that can have significant impacts on companies first and indirectly individuals and society [399,469]. Ensuring that the explanations provided by these systems are not only technically accurate but also aligned with societal norms is a critical prerequisite for responsible adoption [57]. Failure to do so risks not only undermining the credibility of PM and XAI as disciplines but also perpetuating harmful power asymmetries and opaque decision-making practices.

In the sections that follow, we start by overviewing how foundational ethical theories like consequentialism, deontology, and virtue ethics can inform the design and evaluation of explainable PM systems (Section 5.1). We then discuss applied ethics challenges that arise when striving for process transparency in real-world organizational contexts (Section 5.2). Next, we define an XAI risks taxonomy while mapping it to the PM domain, illustrating how various technical and sociotechnical risks can manifest in process analytics workflows (Section 5.3). Finally, we distill a set of ethical principles and risk mitigation strategies that can guide the responsible development and deployment of XAI in PM.

## 5.1 Ethical Theories and Process Mining

Foundational ethical theories provide valuable lenses for analyzing the moral dimensions of XAI systems, including those applied in the PM context. These theories offer frameworks for reasoning about the ethical principles, values, and duties that should guide the design, development, and deployment of possible XAI-enabled PM

systems. By grounding our analysis in these established philosophical traditions, we can surface key ethical considerations and potential tensions that arise when striving for process transparency.

- *Consequentialism* – The findings from Section3.2.2 in Chapter 3 highlight practitioners' emphasis on delivering objective, evidence-based process analyses that can uncover surprising insights and opportunities for improvement. This aligns with the consequentialist focus on maximizing beneficial outcomes, such as process efficiency gains or customer satisfaction. Indeed, consequentialism holds that the moral worth of an action should be judged based on its outcomes or consequences [49, 324, 430]. From this perspective, the ethical evaluation of an explainable PM system would focus on the downstream impacts of the explanations provided, such as their effects on process efficiency, customer satisfaction, or organizational decision-making. A consequentialist approach might prioritize explanations that lead to measurable improvements in process performance metrics or that enable stakeholders to make more informed and beneficial choices. Key considerations would include assessing the accuracy and completeness of explanations as well as anticipating potential unintended consequences [378, 396, 426]. Nevertheless, a narrow focus on outcomes can neglect important moral considerations related to individual rights, autonomy, and dignity [239, 421]. In this spirit, the insights of Chapter 3 also reveal the importance of adopting a diplomatic approach when presenting results that may reveal poor performance, suggesting a need to balance positive impacts with respect for stakeholder dignity and autonomy. Consequentialist evaluations of explainable PM would need to carefully examine not just aggregate metrics but also the distribution of benefits and potential negative side effects across different parties. Yet, considering the lack of user-centric evaluations identified from the SRL in Chapter 2 (aside of virtuous examples such as [404]), these intentions might prove challenging at best, possibly leading to explanations that are not effectively understood or appropriately acted upon.

- *Deontology* – Chapter 3 stresses the value practitioners place on impartiality, empowerment, and collaborative engagement with clients when providing explanations. These strategies resonate with deontological principles [201, 230, 231, 408], that focus on the intrinsic rightness of actions based on moral rules or duties [230, 408]. The core commitment is to ground objective moral principles in the nature of rational agency itself. From a deontological stand-

point, PM explanations would need to respect fundamental human rights like privacy, consent, and non-discrimination, regardless of the aggregated consequences [225,283]. This entails stringent safeguards against misuse of sensitive personal data during process analytics workflows or ensuring that explanations do not inadvertently reveal proprietary information or trade secrets. Deontological considerations would also emphasize the moral obligations of PM practitioners to provide truthful and unbiased explanations, to respect the autonomy of explanation recipients, and to be accountable for the impacts of their systems [201, 253, 372]. Such deontological considerations would further elevate the importance of robust data governance and protection of sensitive information discussed in Chapter 3. Additionally, the lack of standardized benchmarks and evaluation metrics highlighted in Chapter 2 would make challenging to assess whether explanations are truthful, unbiased, and reliable, as demanded by deontological principles.

- *Virtue ethics* – shifts focus from right action to good character, contending that virtues are stable dispositions or traits that reliably lead to human flourishing [29, 210, 298, 437]. While specific virtues may differ across cultures, the basic notion is that character and context are as ethically relevant as actions and their consequences [351]. Neo-Aristotelians argue that truly virtuous agents act for the right reasons and with appropriate emotions, not just in line with moral duties [151, 210, 352]. This approach asks what qualities like honesty, empathy, and wisdom demand in the context of designing and deploying explainable PM systems, e.g. in medical settings with high-stake decisions [305]. Key considerations might include fostering a culture of transparency and openness within PM teams, engaging in ongoing self-reflection and dialogue about the limitations and potential biases of PM algorithms, and proactively seeking out and addressing the concerns of affected stakeholders [121, 282]. The emphasis in Chapter 3 on tailoring explanations to user intentions, organizational culture, and maturity levels aligns with the virtue ethics focus on cultivating context-appropriate qualities and character traits. The insights suggest virtues like empathy, humility (e.g., starting with simple problems), and wisdom (e.g., continuous refinement through feedback) are critical for effective PM explainability. To further complicate that, the underexplored application of XAI in specific domains like healthcare and manufacturing identified in Chapter 2 [150, 398] may hinder the development of context-appropriate virtues.

While these ethical theories offer valuable insights, it can be easily recognized how they can sometimes point in different directions or come into tension with one another in practice [52, 57, 123, 239, 336, 352, 421]. One key debate concerns the relationship between motives, actions, and consequences in determining the moral status of an agent or decision. As seen, deontological theories emphasize the intrinsic rightness of actions based on universal duties, while consequentialist theories focus solely on outcomes. Virtue ethics, meanwhile, stresses the importance of character and moral perception in navigating context-specific challenges [9]. These differences have implications for how XAI systems are designed and evaluated, and for how the decision-making of human agents interacting with these systems is understood and assessed. For instance, the consequentialist aim of optimizing process efficiency might conflict with deontological constraints around data privacy or individual consent. Similarly, the virtue of transparency might need to be balanced against the risks of inappropriately reveal.

The table 5.1 summarizes the key insights and considerations that each ethical approach brings to the domain of PM explainability. It also highlights potential tensions and challenges that may arise when applying these ethical frameworks to real-world PM projects, such as balancing competing priorities, navigating ambiguities, and aligning diverse stakeholder perspectives.

## 5.2 Applied Ethics Challenges in Process Mining

While foundational ethical theories provide valuable frameworks for moral reasoning, translating these abstract principles into responsible practices requires grappling with the concrete realities and constraints of real-world PM projects [304].

This is where the field of applied ethics comes in, developing mid-level principles and context-sensitive guidance to address the moral, political and social implications of technologies like PM in organizational settings [48, 141, 225]. One key challenge is balancing transparency and accountability with the protection of sensitive proprietary information and trade secrets [115, 380]. PM techniques analyze detailed event logs that can reveal insights into an organization's internal processes, resource allocations, and performance bottlenecks, raising risks of inappropriate disclosure or misuse of confidential data. Navigating this tension requires robust data governance frameworks and techniques like data anonymization, aggregation, and differential privacy and careful consideration of trade-offs between data utility and privacy.

Another challenge arises when PM explanations risk perpetuating or amplifying

**Table 5.1:** Ethical Approaches and Their Implications for Process Mining Explainability

| **Ethical Approach** | **Insights for Process Mining** | **Potential Tensions and Challenges** |
|---|---|---|
| Consequentialism | • Focus on delivering objective, evidence-based process analyses that uncover insights for improving efficiency, customer satisfaction, and other beneficial outcomes.<br>• Emphasis on quantifying and maximizing the positive impacts of PM explanations on process performance metrics. | • Need to balance aggregate metrics with potential negative side effects or uneven distribution of benefits across stakeholders.<br>• Risk of oversimplifying process complexities or neglecting individual rights and autonomy in pursuit of optimal consequences.<br>• Challenges in anticipating and mitigating unintended consequences of PM explanations. |
| Deontology | • Importance of impartiality, transparency, and upholding ethical duties when providing PM explanations.<br>• Emphasis on respecting human autonomy, ensuring truthful and unbiased explanations, and protecting sensitive data through robust governance frameworks. | • Potential conflicts between transparency demands and the need to protect proprietary information or trade secrets.<br>• Challenges in navigating ambiguous or conflicting regulatory requirements around explainability across jurisdictions.<br>• Tensions between different moral duties (e.g., privacy vs. accountability) in specific PM use cases. |
| Virtue Ethics | • Need for empathy, humility, and wisdom in tailoring PM explanations to user intentions, organizational culture, and maturity levels.<br>• Emphasis on fostering collaborative platforms, critical self-reflection, and continuous refinement through stakeholder feedback. | • Difficulties in operationalizing and assessing virtuous qualities in PM teams and organizations.<br>• Potential conflicts between cultivating specific virtues and other ethical imperatives (e.g., transparency vs. diplomacy).<br>• Challenges in aligning PM explainability practices with diverse stakeholder values and expectations. |

biases and inequities [314, 375]. The lack of user-centric evaluations considering cognitive and domain factors [404] can exacerbate this, making it difficult to identify and mitigate potential discriminations.

PM explanations also risk being misinterpreted or misused by stakeholders, leading to unjustified actions or decisions. Responsible communication and contextualization is critical, requiring investment in education and training, and establishing clear protocols and accountability mechanisms for how explanations are generated, validated, and acted upon [207]. Finally, navigating complex and conflicting regulatory requirements around PM explainability and transparency across jurisdictions is a major challenge [78, 182, 251]. Balancing legal compliance and stakeholder needs requires ongoing dialogue and collaboration between PM developers, legal experts, and stakeholders to co-create compliant yet meaningful explanations. Proactive engagement with regulators can also help ensure PM standards are appropriately reflected in emerging legal frameworks.

## 5.3 XAI Risks Taxonomy in Process Mining Context

The applied ethics challenges discussed in the previous section lay the groundwork for identifying and categorizing the specific technical and sociotechnical risks associated with deploying XAI techniques in PM. This taxonomy of risks is a natural extension of the broader ethical considerations, translating abstract principles into concrete focal points for assessment and mitigation. While not exhaustive, it provides a structured framework for anticipating and addressing potential harms across key dimensions such as robustness, fairness, privacy, and sociotechnical factors. As PM use cases and XAI techniques evolve, this taxonomy should be refined and expanded based on real-world experiences and research developments.

### 5.3.1 Technical Risks

#### Robustness Risks

Robustness risks in XAI for PM relate to the stability and reliability of explanations when faced with uncertainties, perturbations, or adversarial attacks. These risks can arise from adversarial manipulation of event logs, process model parameters, or explanation methods, leading to inconsistent or misleading interpretations of process models and their predictions.

Attacks on saliency-based explanation methods, such as LIME [402] and SHAP [296], could manipulate or obscure the true importance of activities or resources in process models [436, 509]. For instance, an attacker could perturb event logs to make certain activities appear more or less important in the generated explanations, potentially misleading stakeholders about process bottlenecks or inefficiencies. Potential solutions include robust saliency estimation techniques [10], self-explaining neural networks [24], adversarial training [456, 512], and leveraging adversarial explanations [503]. Counterfactual explanations in PM can be manipulated to deceive users or obscure biases in process models [435, 484]. An attacker could craft adversarial event logs that lead to counterfactual explanations suggesting unrealistic or infeasible process changes. Research focuses on improving counterfactual plausibility [236, 240], incorporating additional constraints [235, 261], and evaluating robustness in specific application domains [327]. Concept-based explanation methods, like TCAV [243], are vulnerable to attacks that can corrupt or misrepresent process-related concepts [61, 170, 434]. An attacker could manipulate the event logs used to define concepts such as "efficient processing" or "customer satisfaction," leading to misleading explanations of the process model's behavior. Potential defenses include detection methods for adversarial examples [170], careful curation and expansion of concept examples [61], and adversarial training [434]. Adversarial data perturbations can affect explanations of process models, reducing their reliability [43]. Techniques to enforce or mitigate the effects of adversarial data perturbations include data poisoning attack strategies or frameworks targeting fairness measures or decision boundaries [314, 440, 510]. [341] examine robustness bias, and [456] propose adversarial training on explanations to improve stability. Explanation-aware backdoors can manipulate explanations to conceal or obfuscate true model behavior [349]. For instance, an attacker could create a backdoored process model that appears to make predictions based on legitimate process features but actually relies on hidden, sensitive attributes. Debugging challenges arise when using post-hoc explanations to diagnose errors in process models [11, 12, 95]. Post-hoc explanations may not always accurately reflect the model's decision-making process, making it difficult to identify and correct errors in process models, such as incorrect process flows or misclassified activities. Lastly, the transferability of adversarial attacks across different explanation methods in PM poses an additional risk [265, 433]. An attack designed to manipulate one explanation method, such as LIME, may also affect other methods, like SHAP or counterfactual explanations, making it more challenging to ensure the overall robustness of the XAI system.

### Fairness Risks

Closely related to robustness are concerns around the fairness and equity implications of XAI in PM. As noted in the applied ethics discussion, PM explanations can perpetuate or amplify discriminatory biases if not carefully designed and audited.

Fairwashing involves the manipulation of explanations to present an unfair PM model as ethical [19, 20]. This deceptive practice distorts fairness metrics, creating a misleading impression of fairness in PM explanations. Biased sampling deceives fairness auditing tools by producing event logs that portray an unfair PM model as unbiased [161, 264]. By carefully selecting a subset of the event log that appears to be fair, biased sampling attacks can manipulate the explanations generated by XAI methods, making it difficult to identify the underlying biases in the PM model. This could lead to unfair resource allocation or process improvements that disadvantage certain groups of process participants. Adversarial poisoning corrupts event logs to induce unfair classification disparities, particularly regarding sensitive attributes [314,440]. By carefully crafting adversarial examples and injecting them into the event logs, adversarial poisoning attacks can manipulate the learned decision boundaries and explanations in PM models, leading to unfair outcomes. The manipulation of post-hoc explanations in PM, as revealed by [320], [108], and [264], involves masking the role of sensitive features and undermining the reliability of remote explainability. By carefully perturbing the event logs or the PM model parameters, an attacker can manipulate the post-hoc explanations generated by XAI methods, hiding the true importance of sensitive features and making the model appear fairer than it actually is. Explanation disparity risks in PM are highlighted by [95] and [41]. These studies suggest that explanation methods themselves can introduce or perpetuate unfairness, even when the underlying PM model is fair. For instance, if the explanation method provides lower-quality explanations for certain subgroups of process participants, it could lead to biased decision-making based on those explanations.

### Privacy & Security Risks

Another key category of risks relates to privacy and security in PM environments. XAI techniques that pinpoint specific process instances, attributes, or individual behaviors as explanatory factors can inadvertently reveal sensitive personal information or enable inferential attacks on the underlying data [391].

For example, an explanation highlighting the role of a particular employee in a process bottleneck could violate that individual's data protection rights if not properly

anonymized. Similarly, an explanation that identifies a rare combination of attribute values as leading to a certain outcome could be reverse engineered to identify specific cases in the event log, compromising data confidentiality [427]. Furthermore, sensitive attributes such as race or sex can be inferred from model explanations, reinforcing the understanding of model explanations as a potent attack surface and a threat to data privacy [110]. Efforts to address these risks include approaches based on Rényi differential privacy (RDP), which ensure robust interpretation through top-k robustness and offer a balance between robustness and computational efficiency [286]. Even if individual-level data is not directly exposed, PM explanations that reveal aggregate patterns of behavior or performance across different process scenarios can still enable adversaries to reconstruct sensitive business logic or proprietary workflows. Explanations that surface the key factors driving process efficiency in a manufacturing plant could be exploited by competitors to infer trade secrets or optimize their own operations [16]. Moreover, disclosing detailed information in explanations can inadvertently provide insight into sensitive intellectual property or trade secrets, allowing competitors or malicious actors to gain an advantage.

### 5.3.2 Sociotechnical Risks

Sociotechnical risks in XAI for PM arise from the complex interactions between the technical aspects of the explanations and the social, organizational, and cultural contexts in which they are developed and used. These risks can lead to unintended consequences, misinterpretations, or even harm to individuals or groups involved in the PM activities. One major sociotechnical risk is the potential for misleading or harmful explanations arising from the interplay of XAI algorithms and human cognitive biases in PM. PM explanations that oversimplify process complexities, highlight spurious correlations, or align with users' prior beliefs can create a false sense of understanding and lead to unjustified actions [411]. Confirmation bias could lead managers to overly fixate on PM explanations that support their existing views while dismissing counterevidence. Anchoring effects could cause decision-makers to rely too heavily on the first explanation provided and fail to seek out alternative perspectives. Overconfidence in XAI could lead to automation bias, where users blindly defer to machine-generated explanations without critical reflection [346, 464]. These cognitive pitfalls can be exacerbated by the inherent uncertainties and ambiguities in many PM event logs, where multiple explanations may be consistent with observed process flows [262, 291, 446].

Another sociotechnical risk is the traceability of explanation design, which refers to the difficulty in identifying the agent responsible for the assumptions underlying an explanation due to the inherent complexity of AI systems and the supply chain related to data lineage and deployment in PM environments [83,97]. This risk can be exacerbated when AI systems are deployed maliciously or manipulated to deceive in PM contexts [495]. The appraisal of explainers also poses a risk, where the perceived expertise and credibility of an explainer may foster unwarranted trust in the explainer's authority and judgments in PM [260, 507]. Explainer's overconfidence can lead to misguided or harmful decisions in process improvement initiatives when explainers overestimate their knowledge or abilities [258, 259, 505]. Heuristics and reception risks can impact the understanding and effectiveness of PM explanations when they are influenced by cognitive biases or heuristics, or misinterpreted by recipients in organizational settings [205, 226, 227, 279, 346, 411, 464]. Argumentative and logical risks, such as circular reasoning [185, 490] and tautology [315], can hinder the transmission of new information and obstruct a deeper comprehension of PM insights within organizations. These risks can be mitigated by designing explanations that are clear, logical, and based on familiar concepts and argumentation style [237, 241, 489].

Underdetermination and overdetermination risks can also pose challenges in developing and presenting PM explanations in organizational contexts. Underdetermination arises when there are several plausible explanations for the same observed phenomena in PM [262, 446], while overdetermination occurs when numerous causes or factors are invoked to explain a single phenomenon, even when they may not all be necessary or directly pertinent [488].

# CHAPTER 6

# A CONCEPTUAL FRAMEWORK FOR EXPLAINABILITY IN PROCESS MINING

Our previous chapters have analyzed explainable AI (XAI) techniques for Process Mining (PM) from multiple perspectives, revealing valuable insights but also complex challenges.

On the technical front, XAI methods overwhelmingly focus on feature relevance explanations like SHAP and LIME over causal or contrastive techniques (Chapter 2). Evaluations emphasize quantitative metrics over qualitative assessments, with minimal emphasis on organizational deployment factors (Chapter 3). A gap persists between proposed XAI solutions and real-world implementation. From an HCI viewpoint, explanations rarely align with mental models of diverse users and neglect incremental querying capabilities (Chapter 3). User interfaces lack personalization, often overwhelming non-technical stakeholders. Linguistic explanations are underutilized compared to visualizations. Practitioners also highlight data and system integration barriers that hinder PM adoption. Legally, extensive documentation mandates for high-risk AI systems under emerging regulations like the EU AI Act place a greater compliance burden on PM providers (Chapter 4). However, individual explainability rights lack concrete standards, discouraging transparency. Legitimate confidentiality protections also limit full model disclosure during audits. Ethically, myriad dilemmas persist around transparency vs privacy, accountability of opaque systems, fairness of predictions, and equitable impacts on diverse users (Chapter 5).

These interlinked challenges underscore the need for a comprehensive framework integrating technical XAI solutions with practical, legal, and ethical considerations in a principled manner.

This chapter presents a comprehensive framework for integrating XAI into PM projects. The framework guides PM practitioners, data scientists, and business stakeholders in developing and implementing explainable PM solutions aligned with their specific goals, requirements, and constraints. Its development is motivated by the need to bridge the gap between XAI technical advancements and the practical challenges of real-world PM implementation, as identified in previous chapters.

We begin by laying the groundwork for the framework in Section 6.1, which details the core components that underpin the integration of XAI in PM. These components, derived from a synthesis of the findings from the previous chapters, include the goals, benefits, costs, risks, and potential negative impacts associated with XAI adoption in PM. The inclusion of these components is motivated by several theoretical perspectives. The "*Goals*" component is informed by goal-setting theory, which posits that clear, specific, and challenging goals can lead to higher performance and motivation [288, 289]. The "*Benefits*," "*Costs/Risks*," and "*Possible Impacts*" components are grounded in the literature on benefit-risk analysis, which emphasizes the need to systematically assess the potential positive and negative consequences of a decision or action [146, 186]. The "*Possible Explanation Targets*" and "*Possible Explanation Types*" components are informed by socio-technical systems theory that recognize the interplay of social and technical factors in organizational systems [166, 460].

With the components established, Section 6.2 introduces the key interaction contexts within the XAI-PM ecosystem. These contexts were identified based on the insights from the empirical study (Chapter 3), which highlighted the importance of considering the diverse needs and perspectives of different stakeholders in PM projects. The selection of these contexts is grounded in stakeholder theory [158, 328], which considers the diverse interests and perspectives of different stakeholders in organizational decision-making processes.

Building upon the components and interaction contexts, Section 6.3 presents a phased approach to implementing the XAI-PM framework. This approach addresses the limitations and research gaps identified in the systematic literature review (Chapter 2) and the empirical study (Chapter 3), which highlighted the need for structured guidance on the integration of XAI in PM projects. This approach is motivated by contingency theory [109], which suggests that the optimal course of action depends on the specific circumstances and context of the situation. By providing step-by-step guidance on context mapping, needs analysis, explanation type selection, technique development, deployment, and monitoring, the framework bridges the gap between the conceptual foundations and practical implementation of XAI in PM.

## 6.1 Components

The proposed framework consists of several key components for understanding and addressing challenges and requirements of XAI adoption in PM contexts (Figure 6.1).

The first-level component is the Actors, which encompasses the key stakeholders and entities involved in the development, deployment, and use of explainable AI solutions in PM. These actors are broadly categorized into three main groups: AI Providers/Deployers, Oversight Bodies, and Process Mining Clients & Affected Stakeholders. For each actor, we consider four sub-components: Goals, Benefits, Costs/Risks, and Possible Negative Impacts. These sub-components are crucial for understanding the specific objectives, advantages, challenges, and potential drawbacks associated with XAI adoption from the perspective of each actor.

The *Goals* component is informed by the practitioner insights in Chapter 3, which emphasized the importance of aligning XAI adoption with specific project goals and stakeholder needs. The *Benefits* and *Costs/Risks* components are grounded also in the technical challenges and limitations identified in the systematic literature review (Chapter 2), as well as the organizational and ethical considerations discussed in Chapters 3 and 5. The *Possible Impacts* component is motivated by the socio-technical risks and ethical implications examined in Chapter 5, highlighting the need to anticipate and mitigate potential negative consequences of XAI deployment in PM contexts. The *Possible Explanation Targets* and *Types* components are informed by the taxonomy of XAI techniques and evaluation approaches presented in Chapter 2, as well as the contextual factors shaping explanation requirements, (Chapters 3 and 4).

To organize and present the various components of the framework in a clear and structured manner, we have introduced a numerical coding system. Each component is assigned a unique code that reflects its hierarchical position within the framework. For example, the goals, benefits, costs/risks, and possible impacts for AI Providers/Deployers are assigned codes such as 1.1, 1.2, 1.3, and 1.4, respectively, where the first digit (1) represents the actor and the second digit represents the specific component. Similarly, the sub-components within each component are assigned a three-part code, such as 1.1.1, 1.2.1, etc., where the third digit represents the specific sub-component.

### 6.1.1 Actors

The actors component of the framework encompasses the key stakeholders and entities involved in the development, deployment, and use of explainable AI solutions in PM contexts.

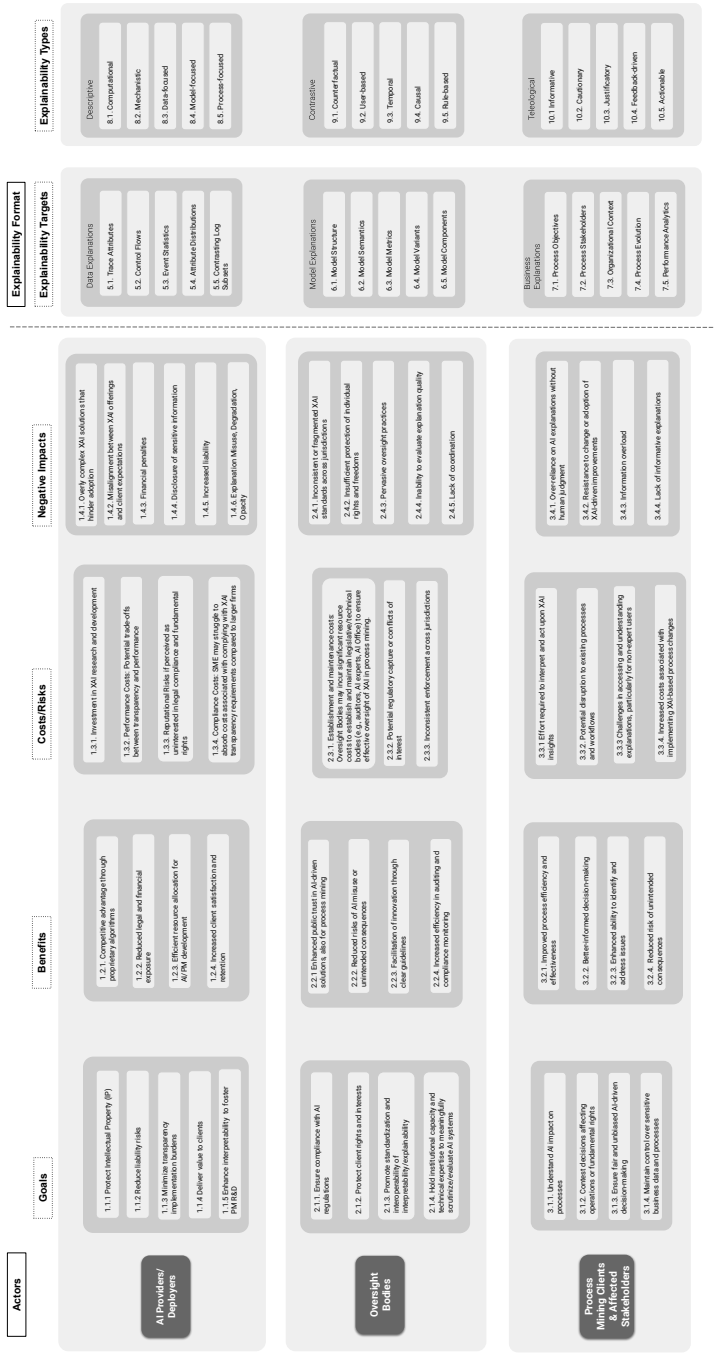


**Figure 6.1:** Visual representation of components.

### AI Providers/Deployers

The goals, benefits, costs/risks, and possible negative impacts for AI Providers/Deployers were derived from the findings of the systematic literature review (SLR) in Chapter 2 and the practitioner study in Chapter 3.

***Goals*** – The goal of *protecting intellectual property* (IP) (PD 1.1.1) is strongly supported by Chapter 4, which highlighted the tensions between transparency and IP/trade secret protection in the context of XAI as also highlighted by [183,338]. This tension was also evident in Chapter 3, where practitioners emphasized the need to balance competitive advantage with legal/ethical considerations. Similarly, the goal of *reducing liability risks* (PD 1.1.2) is grounded in the legal enforceability frictions identified in the SLR [98] and the regulatory challenges discussed in Chapter 4. The goal of *minimizing transparency implementation burdens* (PD 1.1.3) is informed by the technical limitations and scalability challenges of XAI techniques, as discussed in the SLR [66, 208]. These challenges were also echoed in the practitioner study, where participants highlighted the difficulties in generating meaningful explanations for complex PM use cases. Delivering *value to clients* (PD 1.1.4) emerged as a key priority for AI Providers/Deployers in the practitioner study, reflecting the importance of aligning XAI with client expectations and domain-specific needs. Finally, the goal of *enhancing interpretability to foster PM R&D* (PD 1.1.5) is motivated by the potential for XAI to drive innovation and advance the state-of-the-art in process mining, as suggested by the research opportunities identified in the SLR and the practitioner insights on the value of explainability for PM development.

***Benefits*** – The benefits of *competitive advantage through proprietary algorithms* (PD 1.2.1) and *reduced legal and financial exposure* (PD 1.2.2) are firmly grounded in the findings from both the SLR and practitioner study. The SLR highlighted the potential for XAI techniques to enhance the trustworthiness and accountability of AI systems [35], while the practitioner study emphasized the importance of transparent and compliant PM practices for mitigating legal risks. *Efficient resource allocation for AI/PM development* (PD 1.2.3) and *increased client satisfaction and retention* (PD 1.2.4) were identified as key benefits in the practitioner study, reflecting the potential for XAI to optimize PM workflows and improve client outcomes.

***Costs/Risks*** – The costs/risks of *investment in XAI research and development* (PD 1.3.1) and *potential trade-offs between transparency and performance* (PD 1.3.2) are firmly supported by the technical challenges and limitations discussed in the SLR [254, 419]. These challenges were also highlighted in the practitioner study,

where participants noted the resource-intensive nature of implementing robust XAI solutions in PM contexts. *Reputational risks if perceived as uninterested in legal compliance and fundamental rights* (PD 1.3.3) and *compliance costs, particularly for SMEs* (PD 1.3.4), are consistent with the legal enforceability frictions and the need to balance transparency with IP protection, as discussed in Chapter 4 by [115, 183].

***Negative Impacts*** – The possible negative impacts of *overly complex XAI solutions that hinder adoption* (PD 1.4.1) and *misalignment between XAI offerings and client expectations* (PD 1.4.2) are strongly supported by the findings from the practitioner study, which underscored the importance of tailoring explanations to user needs and organizational contexts [404]. *Financial penalties* (PD 1.4.3), *disclosure of sensitive information* (PD 1.4.4), and *increased liability* (PD 1.4.5) are firmly grounded in the legal and ethical risks discussed in Chapter 5 [225]. The additional risks of *explanation misuse*, *explanation degradation*, and *explanation opacity* (PD 1.4.6) – i.e., explanations' information is used by malicious actors to manipulate the AI system and/or the explanations are not actualized and/or understandable by the layperson – were derived from the ethical considerations and sociotechnical challenges analyzed in Chapter 5, particularly the sections on applied ethics challenges and XAI risk taxonomy.

### Oversight Bodies

The goals, benefits, costs/risks, and possible negative impacts for Oversight Bodies were informed by the findings from the SLR, practitioner study, and the analysis of the regulatory landscape in Chapter 4.

***Goals*** – The goal of *ensuring compliance with AI regulations* (OV 2.1.1) is a fundamental responsibility of Oversight Bodies, as highlighted in the context of the EU AI Act in Chapter 4 [131]. *Protecting client rights and interests* (OV 2.1.2) and *promoting standardization and interoperability of interpretability/explainability* (OV 2.1.3) are key objectives for effective oversight (e.g., especially for the AI Office under the EU AI Act) [350]. Building institutional capacity and technical expertise to meaningfully scrutinize/evaluate AI systems (OV 2.1.4) is crucial for effective oversight, as emphasized by [117, 229]. This goal is also supported by the findings from 4, which highlighted the challenges faced by oversight bodies in assessing the quality and appropriateness of PM explanations.

***Benefits*** – The benefits of *enhanced public trust in AI-driven solutions*, also for PM (OV 2.2.1), *reduced risks of AI misuse or unintended consequences* (OV 2.2.2), *facilitation of innovation through clear guidelines* (OV 2.2.3), and *increased efficiency*

*in auditing and compliance monitoring* (OV 2.2.4) are grounded in the findings from the SLR and practitioner study. The SLR underscored the potential for effective oversight to foster public confidence in AI systems, while the practitioner study suggested the importance of clear regulatory guidance for enabling responsible PM innovation.

***Costs/Risks*** – The costs/risks of *establishment and maintenance costs for oversight bodies* (OV 2.3.1) and *potential regulatory capture or conflicts of interest* (OV 2.3.2) are directly informed by the challenges of regulatory capacity and expertise identified by [117, 301]. These challenges were also evident in the practitioner study, where participants noted the resource-intensive nature of effective PM oversight and the need for robust governance mechanisms to ensure impartiality. The risk of *inconsistent enforcement across jurisdictions* (OV 2.3.3) is firmly supported by the legal enforceability frictions discussed by [301] and the regulatory fragmentation analyzed in Chapter 4.

***Negative Impacts*** – The possible negative impacts of *inconsistent or fragmented explainability standards across jurisdictions* (OV 2.4.1), *insufficient protection of individual rights and freedoms* (OV 2.4.2), *pervasive oversight practices* (OV 2.4.3), *inability to evaluate explanation quality* (OV 2.4.4), and *lack of coordination* (OV 2.4.5) are derived from a synthesis of the regulatory challenges and ethical considerations analyzed in Chapters 4 and 5. These impacts are supported by the discussions on the need for harmonized XAI standards [343], the importance of individual rights protection [485], the risks of over-regulation [302], the challenges of evaluating explanation quality [122], and the importance of multi-stakeholder collaboration to relate XAI techniques to policy's requirements [200, 377].

### Process Mining Clients & Affected Stakeholders

The goals, benefits, costs/risks, and possible negative impacts for Process Mining Clients & Affected Stakeholders were informed by the findings from the SLR, practitioner study, and the ethical analysis in Chapter 5.

***Goals*** – The goals of *safeguarding business interests* (CL 3.1.1), *understanding AI impact on processes* (CL 3.1.2), *contesting decisions affecting operations or fundamental rights* (CL 3.1.3), *ensuring fair and unbiased AI-driven decision-making* (CL 3.1.4), and *maintaining control over sensitive business data and processes* (CL 3.1.5) relate to the stakeholder perspectives and ethical considerations discussed in the SLR and practitioner study. The SLR highlighted the importance of explainability for informed decision-making and process optimization, while the practitioner study emphasized the need for transparent and accountable PM practices to ensure stakeholder trust and

engagement.

***Benefits*** – The benefits of *improved process efficiency and effectiveness* (CL 3.2.1), *better-informed decision-making* (CL 3.2.2), *enhanced ability to identify and address issues* (CL 3.2.3), and *reduced risk of unintended consequences* (CL 3.2.4) are strongly supported by the findings from the practitioner study, which underscored the value of PM explanations for process optimization and problem-solving [163]. These benefits are also consistent with the discussions in the SLR on the potential for XAI to enhance the interpretability and actionability of PM insights [496]. The additional benefits of *increased trust and confidence in AI-driven process improvements* (CL 3.2.5), *explanation literacy impact* (CL 3.2.6), *organizational learning impact* (CL 3.2.7), and *accountability and governance impact* (CL 3.2.8) were derived from the ethical analysis in Chapter 5, which emphasized the potential for responsible PM explainability to empower stakeholders and drive positive change. These benefits are further supported by the discussions in the SLR on the importance of user-centric explanations [404] and the need for multi-stakeholder collaboration in PM projects.

***Costs/Risks*** – The costs/risks of *effort required to interpret and act upon XAI insights* (CL 3.3.1), *potential disruption to existing processes and workflows* (CL 3.3.2), *challenges in accessing and understanding explanations, particularly for non-expert users* (CL 3.3.3), and *increased costs associated with implementing XAI-based process changes* (CL 3.3.4) are firmly grounded in the findings from the practitioner study, which highlighted the practical challenges and resource constraints faced by PM clients and stakeholders [143]. These costs/risks are also consistent with the technical limitations and sociotechnical challenges discussed in the SLR [248, 450].

***Negative Impacts*** – The possible negative impacts of *over-reliance on AI explanations without human judgment* (CL 3.4.1), *resistance to change or adoption of XAI-driven improvements* (CL 3.4.2), *information overload* (CL 3.4.3), and *lack of informative explanations* (CL 3.4.4) are derived from the sociotechnical risks and ethical considerations analyzed in Chapter 5, particularly the sections on applied ethics challenges and XAI risk taxonomy [123]. These impacts are further supported by the discussions in the SLR on the importance of human judgment [207], stakeholder engagement [404], and context-sensitive explanation design [150] in PM.

### 6.1.2 Explanations Format

In addition to the actors and their associated goals, benefits, costs/risks, and possible impacts, the proposed framework also considers the format of explanations in terms

of their targets and types. This Explanations Format section is separated from the actors' components to emphasize the importance of designing explanations that align with the specific needs and constraints of PM projects.

The Explainability Targets subsection (Section 6.1.2) focuses on the "*what*" aspect of explanations, identifying the key elements of the PM ecosystem that require explanations. Its inclusion is motivated by Chapter 2, which highlighted the diverse range of PM tasks that can benefit from explanations, such as process discovery, conformance checking, and performance analysis (Section 2.4.2). Additionally, the practitioner study (Chapter 3) emphasized the importance of providing explanations that cater to the specific needs and background knowledge of different stakeholders (Section 3.2.2), which can be facilitated by clearly identifying the targets of explanations.

The Explanation Types subsection (Section 6.1.2) addresses the "*how*" aspect of explanations, exploring the different forms that explanations can take to convey meaningful insights. The motivation for including the Explanation Types subsection is grounded on the limitations found by the SLR (Chapter 2 Section 2.5.1), especially due to the predominance of certain explanation techniques (e.g., LIME, SHAP) and the lack of diversity in the types of explanations provided in current PM research. Moreover, the ethical considerations discussed in Chapter 5 (Section 5.3.2) underscore the importance of providing explanations that go beyond technical transparency and address the broader implications and purposes of PM systems.

### Explainability Targets

Designing explanatory systems for PM requires determining the desired explanation targets. Based on the PM literature [467, 469] as well as the coding we established in the SLR, we identify three key targets for explanations in PM: 1) the *underlying event log data* (XTA 5); 2) the *discovered process model* (XTA 6); and 3) the *organizational business context* (XTA 7).

- **Data explanations** (XTA 5) Explanations are essential for understanding the event log data, which serves as the input for PM. Event logs capture detailed information about process activities and associated data attributes [4, 467]. Key explainability targets for event log data include trace attributes (XTA 5.1), control flows (XTA 5.2), event distributions (XTA 5.3), log subsets (XTA 5.5), and semantic opacity. *Trace attributes* (XTA 5.1) provide essential contextual details about events (e.g., timestamp, activity name, resource, cost), enabling a deeper understanding of their significance [313]. *Control flows* (XTA 5.2) capture

the sequence, concurrency, and dependencies between activities, allowing for the validation of the recorded process execution [101, 303]. *Event statistics* (XTA 5.3) summarize event distributions and resource utilization [140, 187, 477]. *Attribute distributions* (XTA 5.4) reveal temporal correlations and outliers [44]. *Contrasting log subsets* (XTA 5.5) enables comparative analysis between different time periods, user cohorts, or process variants, aiding the identification of performance gaps and improvement opportunities.

- **Model Explanations** (XTA 6) Process models represent the procedural workflow, dependencies, semantics, and behaviors extracted from event logs. Key explainability targets for process models include model structure (XTA 6.1), semantics (XTA 6.2), metrics (XTA 6.3), variants (XTA 6.4), and components (XTA 6.5). *Model structure* (XTA 6.1) encompasses the fundamental elements of a process model, such as nodes (activities, events, states), edges (control flow, message flow), gateways (AND, XOR, complex conditions), subprocesses, and roles (swimlanes, access permissions) [228]. *Model semantics* (XTA 6.2) capture the behavioral aspects of a process, including sequence flows, concurrency, loops, dependencies, and conditions [107, 144, 228, 317]. *Model metrics* (XTA 6.3), such as fitness, precision, generalization, simplicity, and complexity, provide quantitative measures of model quality and understandability [111, 316]. *Model variants* (XTA 6.4) refer to the differences in process models across time, user cohorts, or behavior clusters, enabling comparative analysis and diagnostics [7, 8, 274, 454]. *Model components* (XTA 6.5) include visual elements like highlights, replays, heatmaps, and annotations that enhance the interpretability of process models [8, 56].

- **Business Explanations** (XTA 7) The organizational business context is essential for holistic understanding. Processes are embedded within complex corporate environments with diverse objectives (XTA 7.1), stakeholders (XTA 7.2), policies (XTA 7.3), evolutionary histories (XTA 7.4), and performance metrics (XTA 7.5) [399, 443]. High-level explainability targets for business context include process objectives (XTA 7.1), stakeholders (XTA 7.2), organizational context (XTA 7.3), evolution (XTA 7.4), and performance analytics (XTA 7.5). *Process objectives* (XTA 7.1) encompass the key performance indicators (KPIs) and goal metrics that drive process execution and improvement, such as cost, time, quality, and customer satisfaction. *Process stakeholders* (XTA 7.2) include the various roles and individuals involved in the process, such as owners, operations staff,

support functions, and customers. *Organizational context* (XTA 7.3) refers to the policies, guidelines, and constraints that shape process behavior and decision-making. *Process evolution* (XTA 7.4) captures the changes and adaptations of the process over time, including incremental improvements, continual alignment with changing requirements, and major transformations. *Performance analytics* (XTA 7.5) involve monitoring KPIs, conducting predictive simulations and what-if analyses, and identifying performance drivers.

### Explanation Types

A fundamental consideration in explanatory systems is determining what form explanations should take. Researchers [69, 269] have identified various complementary types of explanations needed to foster multifaceted understanding in AI, that can be declined in PM. At the most basic level, *descriptive explanations* (ETY 8) objectively articulate what the system contains, whether it be event data, process models, or business contexts. However, purely technical transparency is inadequate in isolation [28, 118, 251]. *Contrastive explanations* (ETY 9) are also necessary for conveying meaning through comparative references, like counterfactual tweaking or juxtaposing model-reality differences.

Yet the deepest insights come from *teleological explanations* (ETY 10) that relate technical outputs to higher-level organizational goals, uncertainties, and decision impacts [308, 508].

- **Descriptive explanations** (ETY 8) Descriptive explanations objectively articulate the system's contents, providing technical grounding. *Computational explanations* (ETY 8.1) elucidate technical execution flows and configurations, *mechanistic explanations* (ETY 8.2) detail causal relationships propagating inputs to outputs, *data-focused explanations* (ETY 8.3) convey statistical patterns and stratifications within event logs, *model-focused explanations* (ETY 8.4) detail prediction logics and semantics, and *process-focused explanations* (ETY 8.5) match conceptual framing to users' organizational domains.

- **Contrastive Explanations** (ETY 9) Contrastive explanations highlight comparative differences from reference points, imparting contextual significance. *Counterfactual tweaking* (ETY 9.1) explores output impacts by tweaking inputs through "what-if" interactions, *user-based explanations* (ETY 9.2) spotlight prediction variances across user groups, *temporal explanations* (ETY 9.3) contrast model

versions over time to elucidate drifts, *causal attribution* (ETY 9.4) disambiguates causal pathways by isolating input influence magnitudes, and *rule-based explanations* (ETY 9.5) clarify global model mechanics through simplified local approximations.

- **Teleological explanations** (ETY 10) Teleological transparency relates technical outputs to goals, decisions, and possibilities, providing purpose-driven perspectives. *Informative projections* (ETY 10.1) clarify scenarios linking algorithms to objectives through simulations and feature analysis, *cautionary framing* (ETY 10.2) conveys model limitations and uncertainties, *justificatory explanations* (ETY 10.3) align outputs with objectives and value judgments, *feedback-driven explanations* (ETY 10.4) gather stakeholder input to inform redesigns, and *actionable interventions* (ETY 10.5) operationalize insights to guide decisions.

## 6.2 Interaction Contexts

The XAI-PM framework identifies three primary interaction contexts that shape the requirements and challenges of implementing explainability in PM projects. These macro contexts, informed by Chapter 3 & 4 and grounded in stakeholder theory [158, 328], represent distinct sets of actors and components as illustrated in Figure 6.2. Each interaction context is assigned a numerical code (1 for *Internal Monitoring*, 2 for *Oversight*, and 3 for *External User Interaction*) aligning with the numbering of the corresponding actors (1.1 for AI Providers/Deployers, 2.1 for Oversight Bodies, and 3.1 for Process Mining Clients & Affected Stakeholders) to maintain consistency and clarity throughout the framework.

The *Internal Monitoring* context (1) focuses on the use of XAI techniques by AI Providers and/or Deployers for internal purposes, such as model development, debugging, and performance optimization. This context is crucial for ensuring the quality, reliability, and continuous improvement of PM solutions, as evidenced by the technical challenges and limitations identified in the SLR (Chapter 2).

The *Oversight* context (2) involves the interaction between AI Providers/Deployers and Oversight Bodies, such as regulatory agencies, auditors, and industry standards organizations. This context is essential for ensuring the compliance, transparency, and accountability of PM solutions, as highlighted by the regulatory analysis in Chapter 4, which stressed governance mechanisms and explainability requirements.

The *External User Interaction* context (3) encompasses the engagement between AI Providers/Deployers and the end-users of their PM solutions, such as process own-

ers, domain experts, and business stakeholders. This context is vital for fostering trust, understanding, and effective use of PM solutions, as underscored by the practitioner insights in Chapter 3, which stressed the importance of tailoring explanations to user needs and organizational contexts.

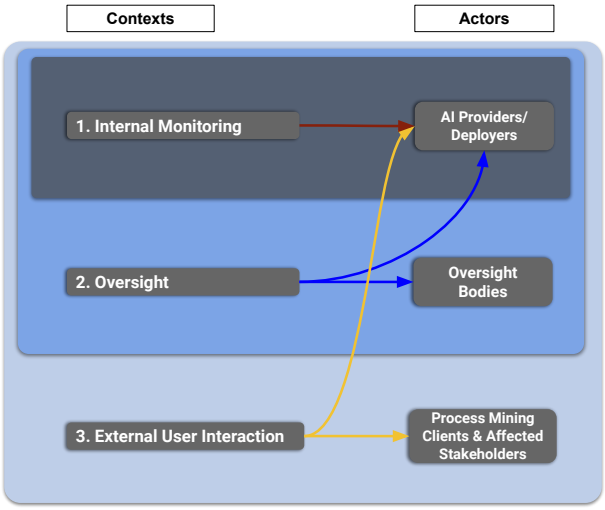


**Figure 6.2:** Visual representation of interaction contexts and related subcontexts. Involved actors are connected by colored arrows. In each context, the main explainer is intended to be the Provider/Deployer.

To further illustrate the interconnections between the various components of the XAI-PM framework, we have developed a comprehensive flowchart (Figure 6.3). This flowchart overviews the relationships between the interaction contexts, subcontexts, actors, goals, benefits, costs/risks, possible negative impacts, and explanations format. The connections between the components are depicted using arrows, indicating the flow of influences and considerations across the framework. These connections are intended to represent tendencies and potential relationships rather than fixed or deterministic links. The actual manifestation of these connections may vary depending on the specific PM project, organizational context, and stakeholder needs.

### 6.2.1 Internal Monitoring

The Internal Monitoring context focuses on the use of XAI techniques by AI Providers and/or Deployers for internal purposes, such as model development, performance optimization, and compliance checking. When selecting explainability targets and explanation types for this context, consider the specific goals and requirements, such

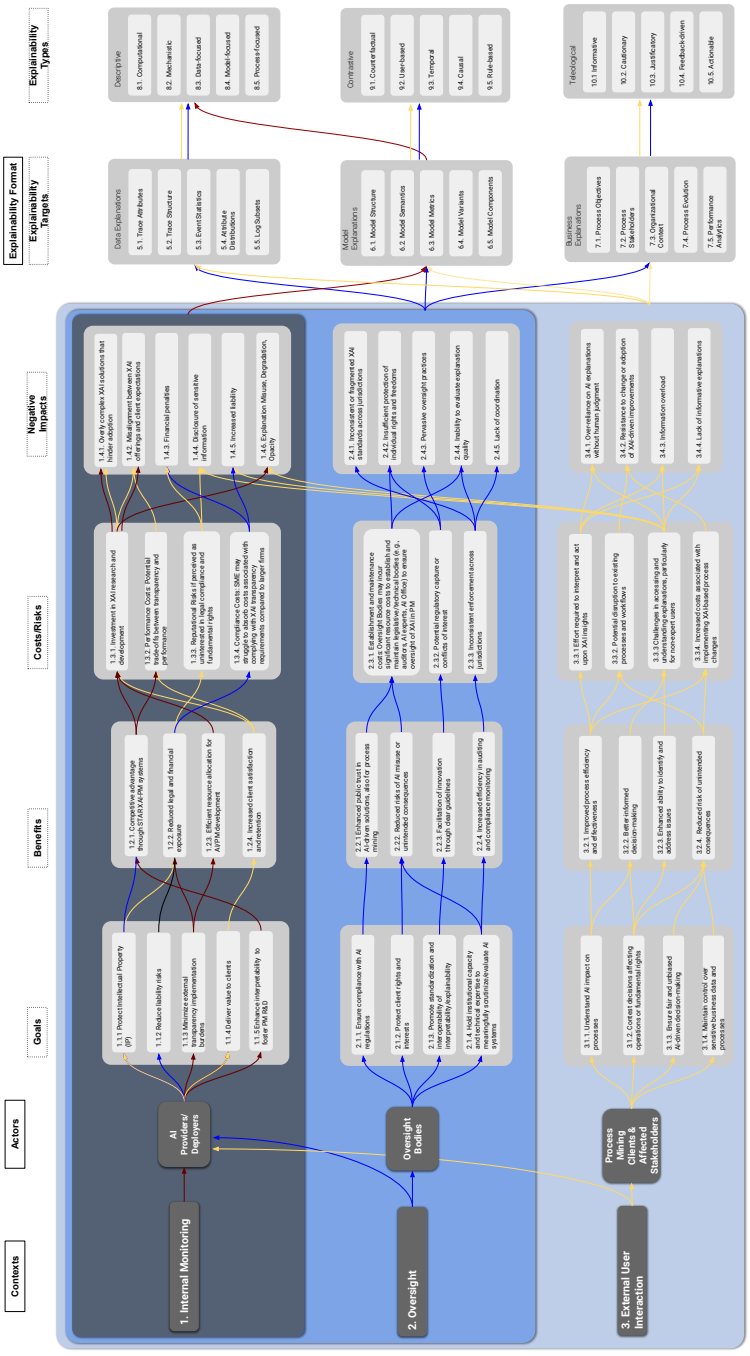


**Figure 6.3:** Flowchart representing the connections between the components of the XAI-PM framework.

as minimizing transparency implementation burdens (PD 1.1.3), delivering value to clients (PD 1.1.4), efficiently allocating resources for AI/PM development (PD 1.2.3), and reducing legal and financial exposure (PD 1.2.2).

The potential benefits in this context include competitive advantage through proprietary algorithms (PD 1.2.1), reduced legal and financial exposure (PD 1.2.2), efficient resource allocation for AI/PM development (PD 1.2.3), and increased client satisfaction and retention (PD 1.2.4). However, there are also costs and risks to consider, such as investment in XAI research and development (PD 1.3.1), potential trade-offs between transparency and performance (PD 1.3.2), reputational risks (PD 1.3.3), and compliance costs (PD 1.3.4). Possible negative impacts include overly complex XAI solutions that hinder adoption (PD 1.4.1), misalignment between XAI offerings and client expectations (PD 1.4.2), financial penalties (PD 1.4.3), disclosure of sensitive information (PD 1.4.4), increased liability (PD 1.4.5), and explanation misuse, degradation, or opacity (PD 1.4.6).

The most appropriate explainability targets and explanation types will depend on factors such as the technical expertise of the users, the desired level of detail, and the specific challenges and limitations of the PM project. For example, model development may benefit from targets related to model structure (XTA 6.1), semantics (XTA 6.2), and metrics (XTA 6.3), while performance optimization may prioritize targets related to event data (XTA 5.1, XTA 5.3, XTA 5.4, XTA 5.5) and performance analytics (XTA 7.5). Computational (ETY 8.1), mechanistic (ETY 8.2), and model-focused (ETY 8.4) explanations may be suitable for model development, while data-focused (ETY 8.3), user-based (ETY 9.2), and actionable (ETY 10.5) explanations may be more relevant for performance optimization and compliance checking.

### 6.2.2 Oversight

The Oversight context involves the interaction between AI Providers/Deployers and Oversight Bodies, such as regulatory agencies, auditors, and industry standards organizations. When selecting explainability targets and explanation types for this context, consider the specific regulatory requirements, auditing processes, and collaborative governance practices relevant to the PM project, such as ensuring compliance with AI regulations (OV 2.1.1), protecting client rights and interests (OV 2.1.2), and promoting standardization and interoperability of interpretability/explainability (OV 2.1.3).

The potential benefits in this context include enhanced public trust in AI-driven solutions (OV 2.2.1), reduced risks of AI misuse or unintended consequences (OV 2.2.2),

facilitation of innovation through clear guidelines (OV 2.2.3), and increased efficiency in auditing and compliance monitoring (OV 2.2.4). However, there are also costs and risks to consider, such as establishment and maintenance costs for oversight bodies (OV 2.3.1), potential regulatory capture or conflicts of interest (OV 2.3.2), and inconsistent enforcement across jurisdictions (OV 2.3.3). Possible negative impacts include inconsistent or fragmented explainability standards across jurisdictions (OV 2.4.1), insufficient protection of individual rights and freedoms (OV 2.4.2), pervasive oversight practices (OV 2.4.3), inability to evaluate explanation quality (OV 2.4.4), and lack of coordination (OV 2.4.5).

The most appropriate explainability targets and explanation types will depend on factors such as the needs of the oversight bodies, the level of transparency required, and the potential risks and negative impacts associated with the PM project. For example, regulatory reporting may focus on targets related to process objectives (XTA 7.1), organizational context (XTA 7.3), and performance analytics (XTA 7.5), while audit and inspection may require a broader range of targets covering event data (XTA 5.1, XTA 5.2, XTA 5.3, XTA 5.4, XTA 5.5), process models (XTA 6.1, XTA 6.2, XTA 6.3, XTA 6.4, XTA 6.5), and business context (XTA 7.1, XTA 7.2, XTA 7.3, XTA 7.4, XTA 7.5). Process-focused (ETY 8.5), cautionary (ETY 10.2), and justificatory (ETY 10.3) explanations may be particularly relevant for regulatory reporting, while a combination of computational (ETY 8.1), data-focused (ETY 8.3), and rule-based (ETY 9.5) explanations may be necessary for audit and inspection.

### 6.2.3 External User Interaction

The External User Interaction context encompasses the engagement between AI Providers/Deployers and the end-users of their PM solutions, such as process owners, domain experts, and business stakeholders. When selecting explainability targets and explanation types for this context, consider the specific needs, preferences, and technical expertise of the end-users, such as understanding the AI impact on processes (CL 3.1.2), contesting decisions affecting operations (CL 3.1.3), ensuring fair and unbiased AI-driven decision-making (CL 3.1.4), and maintaining control over sensitive business data and processes (CL 3.1.5).

The potential benefits in this context include improved process efficiency and effectiveness (CL 3.2.1), better-informed decision-making (CL 3.2.2), enhanced ability to identify and address issues (CL 3.2.3), reduced risk of unintended consequences (CL 3.2.4), increased trust and confidence in AI-driven process improvements (CL 3.2.5),

explanation literacy impact (CL 3.2.6), organizational learning impact (CL 3.2.7), and accountability and governance impact (CL 3.2.8). However, there are also costs and risks to consider, such as effort required to interpret and act upon XAI insights (CL 3.3.1), potential disruption to existing processes and workflows (CL 3.3.2), challenges in accessing and understanding explanations (CL 3.3.3), and increased costs associated with implementing XAI-based process changes (CL 3.3.4). Possible negative impacts include over-reliance on AI explanations without human judgment (CL 3.4.1), resistance to change or adoption of XAI-driven improvements (CL 3.4.2), information overload (CL 3.4.3), and lack of informative explanations (CL 3.4.4).

The most appropriate explainability targets and explanation types will depend on factors such as the users' domain knowledge, the desired level of interaction, and the potential costs and risks associated with transparency and disclosure (CL 3.3.3, CL 3.4.3), contestability and redress (CL 3.3.4, CL 3.4.4), and user empowerment and control (CL 3.3.1, CL 3.3.2, CL 3.4.1, CL 3.4.2). For example, transparency and disclosure may require targets related to event data (XTA 5.1, XTA 5.2, XTA 5.4, XTA 5.5), process objectives (XTA 7.1), and organizational context (XTA 7.3), while contestability and redress may necessitate a broader range of targets covering event data (XTA 5.1, XTA 5.2, XTA 5.3, XTA 5.4, XTA 5.5), process models (XTA 6.1, XTA 6.2, XTA 6.3, XTA 6.4, XTA 6.5), and business context (XTA 7.1, XTA 7.2, XTA 7.3, XTA 7.4, XTA 7.5). Data-focused (ETY 8.3), process-focused (ETY 8.5), and informative (ETY 10.1) explanations may be suitable for transparency and disclosure, while counterfactual (ETY 9.1), user-based (ETY 9.2), and actionable (ETY 10.5) explanations may be more relevant for contestability and redress.

To further illustrate the interconnections between the various components of the XAI-PM framework, Table 6.1 summarizes the key aspects of each interaction context, including the corresponding numerical codes of the associated components.

## 6.3 Phased Approach

Bridging the gap between XAI advances and practical organizational implementation requires an approach spanning precursory needs gathering, tailored explainability techniques and organizational integration support. Indeed, the SLR in Chapter 2.5 revealed that prevailing PM systems often provide outputs lacking explanatory interfaces, failing to clarify insights for business stakeholders and hindering adoption. The interviews findings in Chapter 3.2.2 further highlighted the challenges in aligning XAI offerings with client expectations and domain-specific needs, emphasizing user-centric design and evaluation approaches.

**Table 6.1:** Interaction Contexts and corresponding numerical codes of components

| Context | Involved Codes |
|---|---|
| *Internal Monitoring* (Providers and Deployers) | Goals: 1.1.2, 1.1.3, 1.1.4<br>Benefits: 1.2.1, 1.2.2, 1.2.3, 1.2.4<br>Costs/Risks: 1.3.1, 1.3.2, 1.3.3, 1.3.4<br>Negative Impacts: 1.4.1, 1.4.2, 1.4.3, 1.4.4, 1.4.5, 1.4.6, 1.4.7, 1.4.8<br>Explainability Targets: 5.1, 5.3, 5.4, 5.5, 6.1, 6.2, 6.3, 6.4, 6.5, 7.5<br>Explanation Types: 8.1, 8.2, 8.3, 8.4, 9.2, 10.5 |
| *Oversight* (Providers/Deployers and Oversight Bodies) | Goals: 2.1.1, 2.1.2, 2.1.3<br>Benefits: 2.2.1, 2.2.2, 2.2.3, 2.2.4<br>Costs/Risks: 2.3.1, 2.3.2, 2.3.3<br>Negative Impacts: 2.4.1, 2.4.2, 2.4.3, 2.4.4, 2.4.5<br>Explainability Targets: 5.1, 5.2, 5.3, 5.4, 5.5, 6.1, 6.2, 6.3, 6.4, 6.5, 7.1, 7.2, 7.3, 7.4, 7.5<br>Explanation Types: 8.1, 8.3, 8.5, 9.5, 10.2, 10.3 |
| *External User Interaction* (Affected Individuals and Providers/Deployers) | Goals: 3.1.2, 3.1.3, 3.1.4, 3.1.5<br>Benefits: 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.2.5, 3.2.6, 3.2.7, 3.2.8<br>Costs/Risks: 3.3.1, 3.3.2, 3.3.3, 3.3.4<br>Negative Impacts: 3.4.1, 3.4.2, 3.4.3, 3.4.4<br>Explainability Targets: 5.1, 5.2, 5.3, 5.4, 5.5, 6.1, 6.2, 6.3, 6.4, 6.5, 7.1, 7.2, 7.3, 7.4, 7.5<br>Explanation Types: 8.3, 8.5, 9.1, 9.2, 10.1, 10.5 |

To tackle these barriers and help to identify suitable components and interaction contexts, we propose a phased procedure (Figure 6.4) to translate high-level user requirements into tailored XAI-PM capabilities. The preliminary gathering phase (Section 6.3.1) draws upon the practitioner insights from Chapter 3, outlining methodologies to elicit requirements. Sections 6.3.2 and 6.3.3 entail analyzing a spectrum of needs to shape suitable explanation format considering stakeholders' perspectives and organizational factors of Chapters 3 and 4. The technical tool development phase (Section 6.3.5) refers to XAI techniques and evaluation approaches of Chapter 2, providing directions on developing aligned XAI tools while incorporating ethical considerations of Chapter 5. Finally, the organizational integration phase (Section 6.3.6) draws upon the regulatory analysis in Chapter 4 and the ethical principles of Chapter 5, outlining change management procedures that leverage governance within evolving organizational constraints.

### 6.3.1 Step 0: Explainability Requirements Elicitation

A preliminary departing stage entails being able to gather broad information regarding stakeholders' explanatory contexts and needs. Techniques to gather this information vary considerably based on factors like team size, budget, and client openness. For instance, a consultant with limited access may rely more on small-scale proxy inquiries,

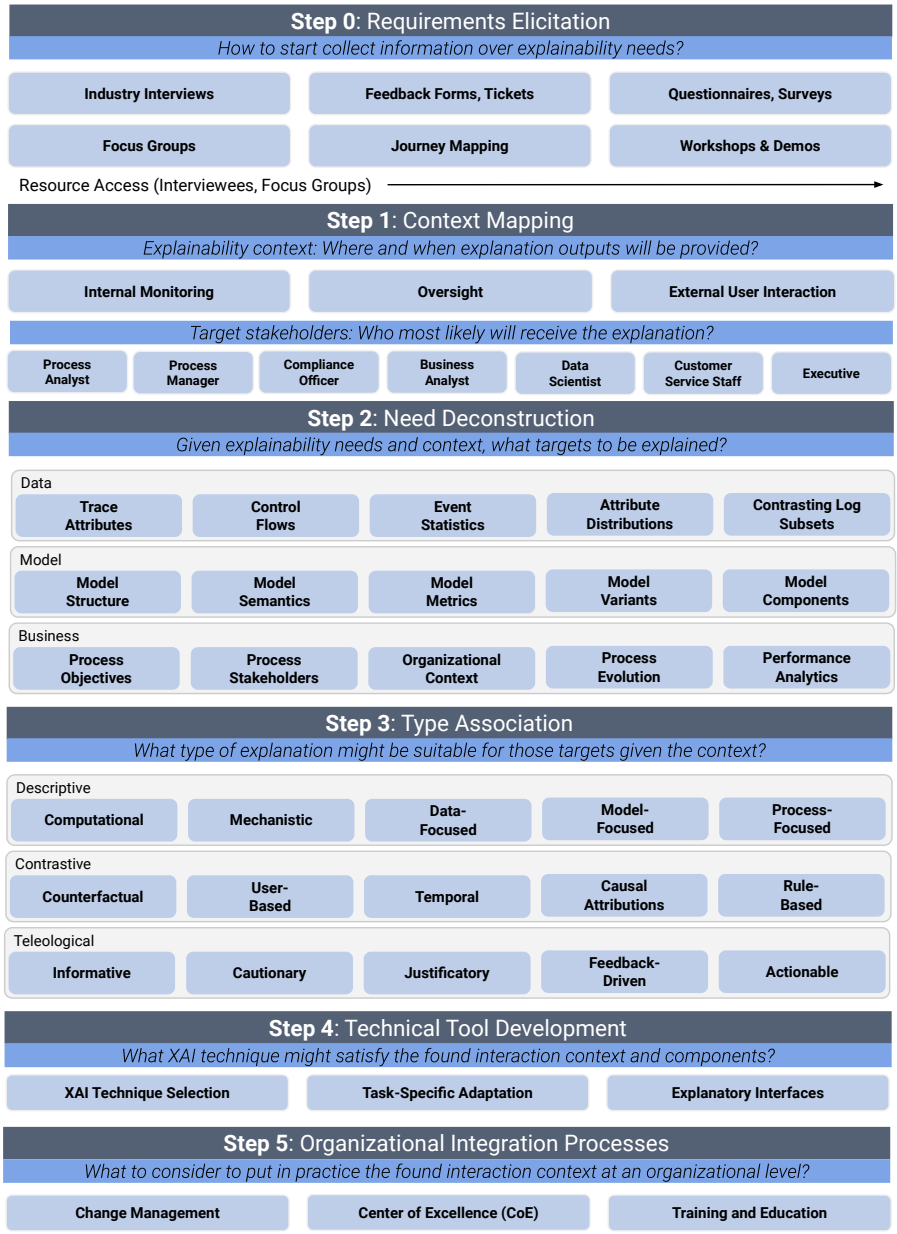


**Figure 6.4:** Procedure of the phased approach

while an internal analytics team can leverage extensive surveys and focus groups. Specifically, explanatory needs analysis encompasses a spectrum of qualitative and quantitative methods. Table 6.2 presents a comparison of elicitation methods both for scenarios with limited and extensive stakeholder access.

| **Limited Access** | **Extensive Access** |
|---|---|
| Industry proxy inquiries | Focus groups for qualitative insights |
| Feedback forms or tickets<br>(5-10 key questions) | Surveys to quantify needs<br>(screen for statistical significance) |
| Needs analysis templates<br>(guidance + examples provided) | Persona synthesis into archetypes<br>(reflecting top clusters) |
| Expectation setting<br>(transparency on input limitations) | Journey mapping showing pain points<br>(visualizing interactions needing explanations) |
| Simplified self-service<br>(quick turnaround polls/forms) | Design workshops and demos<br>(conceptualizing solutions collaboratively) |

**Table 6.2:** Requirements Elicitation Strategies

In scenarios with limited stakeholder access, pragmatic approaches focus on gathering insights from a few priority personas rather than quantitative data from many users. Explanations tailored to addressing recurring top needs of key roles will provide the most value, with continuous input enabling incremental refinements. Conversely, in scenarios with extensive stakeholder access, such as large collaborative projects, more thorough techniques facilitate broad quantitative analysis.

The explainability requirements elicitation step is crucial for understanding the specific needs and expectations of stakeholders across the three interaction contexts outlined in Section 6.2. By gathering information about explanatory needs and intentions, practitioners can identify the relevant actors (AI Providers/Deployers, Oversight Bodies, and Process Mining Clients & Affected Stakeholders) and their goals, as described in Section 6.1. This step helps validate the theorized benefits, risks, and impacts of XAI adoption for each interaction context.

### 6.3.2 Step 1: Context Mapping

After the initial elicitation of explanatory needs in Step 0, the next phase involves translating those insights into practical solutions that fit organizational realities. This step directly addresses the interaction contexts described in Section 6.2, as it involves specifying the capabilities, timing, and target recipients of explanations for each context. By systematically mapping target explanations across contextual dimensions (stakeholders, policies, objectives, risks), this process helps realize the specific goals,

benefits, costs/risks, and possible impacts for each interaction context and the associated actors.

For instance, in the Internal Monitoring context (Section 6.2.1), context mapping can help AI Providers/Deployers achieve the goal of protecting intellectual property (PD 1.1.1) while minimizing transparency implementation burdens (PD 1.1.3). In the Oversight context (Section 6.2.2), this step can assist Oversight Bodies in ensuring compliance with AI regulations (OV 2.1.1) and protecting client rights and interests (OV 2.1.2). The situational relevance, utility, and optimal presentation of explanations can vary significantly across use cases and intended recipients – e.g., real-time interactive explanations for frontline staff in the External User Interaction context (Section 6.2.3) may have different requirements than periodic summary reports for leadership in the Internal Monitoring context (Chapter 3, [143]).

The explanatory personas and archetypes developed earlier provide a foundation for scoping contexts suited to each user segment. Personas capture stakeholders' needs and constraints. For example, a "Data Scientist" persona in the Internal Monitoring context may prefer transparency into technical model architecture, while a "Business Analyst" persona in the External User Interaction context may seek high-level goal alignment insights (Chapter 3, [404]).

The framework's spectrum of explanatory contexts further grounds the analysis. Factors such as organizational workflows, decision types, risk severity, data sensitivity, information dispersion, and governance policies all help constrain explanatory contexts: in particular, we refer to contextual needs of users as reported in Chapter 3.2.2 as well as organizational considerations over compliance of Chapter 4 and sociotechnical risks in Chapter 5.

### 6.3.3 Step 2: Need Deconstruction

With explanatory contexts and target stakeholders identified, the next step is to map the elicited user needs to the specific explainability targets outlined in Section 6.1.2. This need deconstruction process involves analyzing the transparency requirements of each interaction context and selecting the most relevant targets from the three main categories: data explanations (XTA 5), model explanations (XTA 6), and business explanations (XTA 7).

When deconstructing the needs, consider the following guidelines for each interaction context:

1. **Internal Monitoring context:**

- Focus on targets that provide detailed insights into the event log data, such as trace attributes (XTA 5.1), trace structure (XTA 5.2), and event statistics (XTA 5.3). These targets will help process analysts and data scientists identify underlying patterns/anomalies in the process execution data. In this spirit, [140], [477], and [187] highlight the importance of event statistics and attribute distributions for process understanding and anomaly detection (Section 2.4.1).
- Select model-related targets that enable performance optimization and debugging, such as model semantics (XTA 6.2), model variants (XTA 6.4), and model components (XTA 6.5). These targets will support the identification of bottlenecks, inefficiencies, and improvement opportunities in the process models (e.g., variants in enhancing process model comprehension and analysis, as seen in studies like [228], [107], and [8] in Chapter 2.4.2).

2. **Oversight context:**

   - Prioritize targets that facilitate compliance checking and auditing, such as trace structure (XTA 5.2), attribute distributions (XTA 5.4), and log subsets (XTA 5.5). These targets will help auditors and regulators verify the conformance of the process execution data with the prescribed rules and standards, as seen in [303] and [101] stressing over control flow and trace attributes for conformance checking (Chapter 2.4.3).
   - Focus on model-related targets that enable transparency and interpretability, such as model structure (XTA 6.1), model semantics (XTA 6.2), and model metrics (XTA 6.3). These targets will support the assessment of the process models' quality, fairness, and adherence to regulatory requirements [304].

3. **External User Interaction context:**

   - Select targets that align with the organizational objectives and stakeholder requirements, such as process objectives (XTA 7.1), process stakeholders (XTA 7.2), and organizational context (XTA 7.3). These targets will help process owners, domain experts, and business users understand how the AI-driven insights relate to their specific goals and constraints [399, 443].
   - Focus on targets that enable actionable decision-making and continuous improvement, such as process evolution (XTA 7.4) and performance analytics (XTA 7.5). These targets will support users in identifying trends,

making data-driven decisions, and adapting to changing business needs, as reproved by [7,274] over the role of process variants and performance indicators in supporting process optimization and decision-making in Chapter 2.4.2).

### 6.3.4 Step 3: Type Association

With transparency targets and users determined via contextual mapping, optimal explanations get systematically associated to each granular need from among *descriptive* (i.e., factually communicating system compositions) (ETY 8), *contrastive* (highlighting comparative reference differences) (ETY 9), and *teleological* (relating outputs to higher-level goals) (ETY 10) (Chapter 2, [69]). This step is closely linked to the Possible Explanation Types component of the framework, as described in Section 6.1.2.

For instance, contrastive explanations help validate model behaviors by showing deviations from expectations, which is particularly relevant in the Oversight context (Section 6.2.2) for ensuring compliance with AI regulations (OV 2.1.1). Cautionary teleological explanations build trust by qualifying predictions, supporting the goal of safeguarding business interests (CL 3.1.1) in the External User Interaction context (Section 6.2.3).

Tailoring explanatory combinations fitted to personas' needs enables multifaceted transparency, as highlighted in the practitioner study (Chapter 3, [404]). To guide association, facets like intentions (purely explaining internals, externally validating behaviors, linking outputs to goals) should get mapped. Additionally, mapping targets of explanation (computations, data, models or business environments); time references (current states, temporal drifts, future possibilities); comparative needs (facts alone or relative contrasts); uncertainty factors (limitations relevance); participation elements (stakeholder inputs); interactivity (static or interactive); and localization (granular or global) facilitates aligning techniques to needs and constraints.

This process helps validate the appropriateness and effectiveness of the theorized explanation types for each interaction context and actor. For instance, in the Oversight context, type association can help AI Providers/Deployers and Oversight Bodies identify the most suitable explanation types, such as process-focused (ETY 8.5), cautionary (ETY 10.2), and justificatory (ETY 10.3) explanations, to support the goals of reducing liability risks (PD 1.1.2) and ensuring compliance with AI regulations (OV 2.1.1).

### 6.3.5 Step 4: Technical Tool Development

With requirements systematically elicited from stakeholders, the next phase entails translating needs into functioning XAI tools fitted for organizational processes and constraints. This process helps validate the feasibility and appropriateness of the selected XAI techniques for each interaction context and actor, ensuring alignment with the goals, benefits, costs/risks, and possible impacts outlined in Section 6.1.

Chapter 2.5 revealed a heavy reliance on certain explanation techniques like LIME and SHAP in current XAI-PM research, with fewer studies exploring causal or contrastive methods [122, 163, 379]. In the Explanation Types section (Section 6.1.2) is shown how different XAI techniques have complementary strengths and weaknesses that suit them to certain tasks and stakeholders [35]. This underscores the need for guidance on selecting appropriate XAI methods that align with the nuanced explainability needs of diverse PM use cases across the three interaction contexts (Section 6.2). To address this gap, we now outline trade-offs during XAI technique selection and configuration for PM systems.

For process prediction and recommendation tasks, which are particularly relevant in the Internal Monitoring (Section 6.2.1) and External User Interaction (Section 6.2.3) contexts, post-hoc explanation methods like SHAP and LIME have been widely used to provide feature importance scores and local explanations for black-box models [164,448]. Galanti et al. used SHAP to explain the predictions of an LSTM model for cycle time estimation, identifying the key features driving the predictions [163]. However, the survey also highlighted the need for more causal and counterfactual explanations in process prediction tasks [8, 66, 339]. Hinkka et al. proposed a counterfactual explanation method based on decision trees, while Huang et al. introduced LORELEY, a counterfactual explanation technique tailored for predictive business process monitoring [208]. When adapting these techniques for process prediction tasks, consider preprocessing the event log data to extract relevant features, selecting appropriate perturbation strategies for generating counterfactuals, and evaluating the stability and fidelity of the explanations using metrics like the local fidelity score or the causal local explanation score [26, 480].

For conformance checking and anomaly detection tasks, the survey identified techniques based on alignments and rule-based approaches [80, 334]. To enhance the interpretability of alignment-based explanations, consider using multi-perspective alignments, developing visual explanations, and applying techniques like inductive logic programming or decision tree learning to extract interpretable rules from the

alignments [275]. For anomaly detection, clustering techniques can be used to identify unusual patterns. To explain the detected anomalies, consider generating prototypes or exemplars for each anomaly cluster, using feature importance methods to identify the key attributes contributing to the anomaly scores, and developing contrastive explanations to highlight the differences between anomalous and normal instances [80, 339].

For process discovery tasks, the survey identified the use of directly-follows graphs and process trees as interpretable representations of process models [189, 207]. To further improve the interpretability of discovered process models, consider applying techniques like abstraction or modularization to simplify complex models, incorporating domain knowledge and business rules into the discovery algorithms [142, 150].

For process enhancement tasks, which are particularly important in the External User Interaction context (Section 6.2.3), the survey mentioned the use of simulation and optimization techniques to identify improvement opportunities [137, 339]. To provide interpretable explanations for the suggested enhancements, consider using sensitivity analysis to identify the key factors influencing process performance, applying multi-objective optimization techniques to explore trade-offs between different performance measures, and generating what-if scenarios to demonstrate the potential impact of proposed process changes [271, 398].

### Explanatory Interfaces

In addition to selecting appropriate XAI techniques, it is crucial to consider the overall design and integration of the explanation interfaces into the PM system workflow [143, 163, 339, 404]. Indeed, crafting explanatory interfaces that effectively convey insights and align with diverse users' mental models proves equally crucial for utility across all three interaction contexts (Section 6.2). To intuitively convey explanatory information, interfaces should reflect users' mental models, leveraging appropriate vocabulary, visual metaphors, interaction paradigms, and workflow alignments [297]. Incorporating domain familiarity into explanations enhances understanding by leveraging users' existing conceptual framework [30,33]. Specific techniques to align PM-XAI interfaces with mental models include using natural language and terminology familiar to the user's domain [506], employing visual metaphors and icons that map to the user's context [144, 318], designing interactions that match workflows [30], reducing cognitive load by limiting unnecessary interactivity [51, 318], and spatially grouping related information [144, 145]. Additionally, leveraging stake-

holder personas allows tailoring information scent and appeal of explanations to align with user needs. Periodic contextual inquiries help ensure continued alignment with evolving user mental models.

Personalized and adaptive explanation systems using natural language generation (NLG) techniques can enhance the relevance of PM insights for individual users. NLG techniques applicable for explanatory PM systems include two macro approaches: transforming processes into textual narratives (*process-to-text*), and morphing textual rules and descriptions into visual process representations (*text-to-process*). The *process-to-text* strategy involves crafting linguistically lucid narratives that encapsulate control flow features, KPI behavior, temporal dynamics, and other salient attributes of the process. Template-based methods [276,406,442] are adept at converting structured process model representations into lucid textual descriptions, while end-to-end neural networks [68,177,483] generate language descriptions directly from process data without the constraints of templates. Conversely, *text-to-process* methods prioritize the transformation of textual process descriptions into visually coherent representations like control-flow diagrams, using techniques such as dependency parsing and semantic role labeling [418,455,466], automatic extraction of declarative process models [465], and formal reasoning based on natural language descriptions [417]. Natural language interfaces [45,249] and queries [160,187,506] have also been investigated to facilitate user interaction with process data.

### 6.3.6 Step 5: Organizational Integration Processes

As last considerations, integrating PM-XAI systems within organizations requires a multifaceted approach encompassing change management, training, and enhanced communication to foster organization-wide transparency and accountability [399, 469]. Change management begins with comprehensive stakeholder impact assessments, analyzing potential changes in workflows, roles, and organizational structure [32]. A balanced transition plan with a gradual introduction of new tools and the use of change champions can mitigate risks associated with sudden operational changes [204, 255, 323]. Integrating PM and *organizational change management* (OCM) aligns project management practices with business strategies and enhances project success when combined with benefits management practices [39, 204, 445]. Education and training are crucial for developing organizational expertise in PM-XAI systems. Diverse training curricula should cater to different audiences, with immersive workshops for analysts and technical staff and gamification elements to enhance

engagement [116, 287]. The integration process aligns with the trend of establishing *Centers of Excellence* (CoE) in PM, which centralize expertise, standardize practices, and drive the strategic implementation of PM tools across organizational domains (e.g., *athenahealth*, *Telekom*, etc in [399]).

## 6.4 Framework Scope & Directions

The proposed framework provides a comprehensive approach to accelerate responsible progress in XAI systems in PM. It uniquely combines ethical, legal, practical, and technical considerations. Yet, conducting real-world pilot projects across various PM applications would offer valuable insights and opportunities for customization. Creating a library of adaptable templates, tools, and best practice resources could also make it easier to tailor the framework to specific contexts.

To support the practical implementation of the phased approach outlined in this framework, we have developed a comprehensive self-assessment tool that guides users through the key steps and considerations for integrating XAI into their PM practices. The self-assessment tool, presented in Appendix C, is firmly grounded in the conceptual foundations of the framework and provides a structured approach to operationalizing the guidelines and recommendations discussed in this chapter.

By completing the self-assessment, users can identify the specific requirements and challenges of their PM projects and receive tailored guidance on how to implement XAI solutions that meet their unique needs and priorities.

# Chapter 7

# Case Study: Enhancing Explainability in Hospital Process Mining

In the preceding chapters, we have explored the multifaceted challenges and opportunities surrounding explainable AI (XAI) in the context of Process Mining (PM). Having established this conceptual foundation, we now present a hypothetical case study that illustrates how the explainability framework proposed in Chapter 6 can be operationalized in a healthcare setting.

The case study draws inspiration from inVerbis, a Spanish PM company that was founded at CiTIUS of University of Santiago de Compostela. Research at CiTIUS often informs PM techniques and approaches to strengthen different domain solutions from inVerbis, also in healthcare settings to optimize efficiency and comprehension of clinical processes [150,219,220]. The hospital network in question has been exploring the potential of PM techniques to optimize their clinical processes and improve patient outcomes. One area of particular interest is the management of patients with heart valve diseases, such as aortic stenosis, who require complex surgical interventions. The hospital recognizes the need for a more data-driven approach to understanding and improving this critical process.

## 7.1 Process Overview and Challenges

The surgical procedure for heart valve disease patients is a complex, multi-stage process that involves numerous stakeholders, decision points, and potential complications.

The process typically begins with a patient presenting symptoms such as chest pain, shortness of breath, or fatigue, which prompt a referral to a cardiologist. The cardiologist have a first *Medical-Surgical Session* (SMQ) to assess if the patient needs to (i.) undergo a *open-heart surgery*; or (ii.) a *TAVI* (Transcatheter aortic valve implantation - a common procedure to improve blood flow); or (iii.) being dismissed with no further actions required (*conservative management*). If the latter option is discarded, then the cardiologist prescribes a series of diagnostic tests, such as echocardiograms, cardiac catheterizations, and CT scans, to assess the severity of the valve disease and determine the most appropriate treatment option among surgery or TAVI.

One of the key challenges in managing patients with severe symptomatic aortic stenosis is minimizing the delay to intervention, as these patients are at high risk of adverse events while waiting for treatment - e.g., while waiting for unfixed period of times in what can be defined an *elective pathway*, compared to the *emergency pathway* that require a urgent surgical intervention.

In cases where surgical intervention is deemed necessary, the patient is referred to a cardiac surgeon for further evaluation. The surgeon reviews the patient's medical history, imaging results, and other relevant factors to determine the specific type of valve repair or replacement procedure that is required. This decision is often made in consultation with a team of specialists, including anesthesiologists, intensivists, and nursing staff, to ensure that all aspects of the patient's care are considered.

Once the surgical plan is established, the patient in this elective pathway undergoes a series of pre-operative assessments and preparations, which may include blood tests, chest X-rays, and consultations with other specialists to optimize their health status prior to surgery. The actual surgical procedure involves a complex sequence of steps, including anesthesia, cardiopulmonary bypass, valve repair (TAVI) or replacement (Open-heart surgery), and post-operative monitoring and care. Throughout this process, challenges and potential sources of variability can impact patient outcomes and operational efficiency.

Some of the key challenges include:

1. *Minimizing decision time*: Ensuring that patients receive timely and appropriate care is critical in the management of heart valve disease, as delays in treatment can lead to worsening symptoms, complications, and adverse outcomes. However, the complex nature of the diagnostic and treatment process, as well as resource constraints and competing priorities, can sometimes result in longer-than-optimal decision times.

2. *Managing patient complications*: Patients with heart valve disease often have complex medical histories and comorbidities that can impact their surgical risk and post-operative recovery. Identifying and managing these complications requires careful coordination among multiple specialists and ongoing monitoring and adjustment of the care plan.

3. *Ensuring compliance with service level agreements* (SLAs): The hospital has established internal SLAs that define target timeframes for key process milestones, such as the maximum allowable wait time between diagnosis and surgery. Ensuring compliance with these SLAs is important for maintaining quality of care and patient satisfaction, but can be challenging given the complexity and variability of the process.

4. *Extracting actionable insights from process data*: The heart valve disease surgical procedure generates a large volume of data, including clinical notes, imaging results, lab tests, and operational metrics. However, much of this data is currently captured in unstructured or siloed formats, making it difficult to analyze and extract meaningful insights that can drive process improvement.

To support this data-driven approach, the hospital has been collecting and storing a wide range of data related to the heart valve disease surgical procedure. This includes approximately 700 variables capturing patient demographics, clinical characteristics, diagnostic test results, and treatment outcomes, as well as 27 declarative constraints reflecting the temporal and logical dependencies between process steps. These constraints were elicited through discussions with clinical experts.

The heart valve disease surgical procedure is illustrated in Figure 7.1, which depicts the key decision points, diagnostic tests, treatment pathways, and potential complications along the patient journey.

## 7.2 Phased Approach to Explainability

To address the challenges and opportunities identified in the heart valve disease surgical procedure, the PM experts, as AI Providers/Deployers (1.), would collaborate with the hospital, as Process Mining Clients & Affected Stakeholders (3.), to apply the phased approach introduced in Chapter 6, focusing on the External User Interaction context (Section 6.2.3). This section details how each step of the framework could be adapted to the specific needs and constraints of the healthcare process.

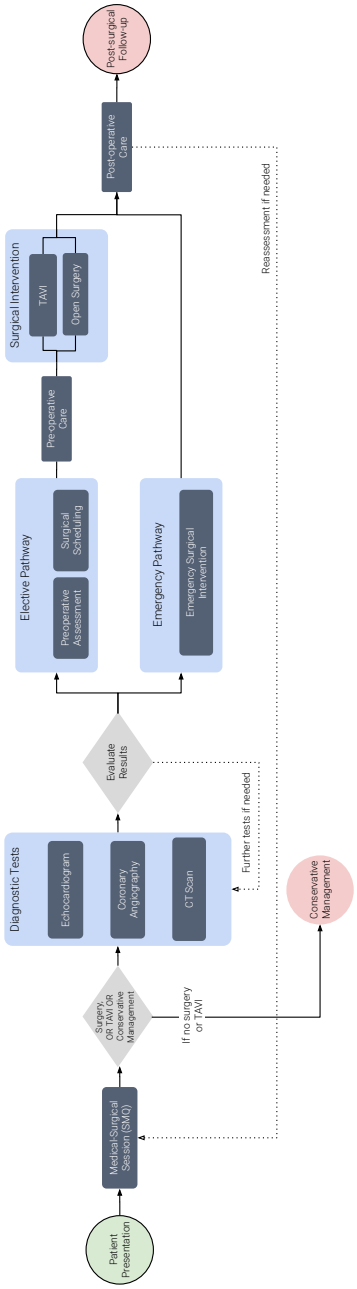


**Figure 7.1:** Process map of the heart valve disease surgical procedure, showcasing the elective and emergency pathways, key decision points (SMQ), diagnostic tests, treatment options (TAVI, surgery, conservative management), and potential complications.

### 7.2.1 Step 0: Explainability Requirements Elicitation

The first step in the framework would involve eliciting the explainability requirements from the key stakeholders involved in the process. The PM experts would conduct a series of workshops and interviews with clinicians, including cardiologists, cardiac surgeons, anesthesiologists, and nursing staff to understand their specific needs and expectations for explainable PM insights [163]. By focusing on the goals of understanding the AI impact on processes (CL 3.1.2) and ensuring fair and unbiased AI-driven decision-making (CL 3.1.4), the PM experts can ensure that the explainability requirements align with the needs of the hospital stakeholders in the External User Interaction context. Clinicians may also express the need to understand the factors influencing the time from the last SMQ to the intervention date, as this can help identify bottlenecks and opportunities for process improvement, aligning with the goal of improving process efficiency and effectiveness (CL 3.2.1).

Through these discussions, several key explainability requirements could emerge:

1. *Delay to intervention*: Clinicians might emphasize the importance of understanding the factors that influence the time from the last SMQ to the intervention date. This requirement aligns with the goal of understanding the AI impact on processes (CL 3.1.2) and the benefit of improved process efficiency and effectiveness (CL 3.2.1).

2. *Events while waiting for intervention*: Experts might express a need to quantify and analyze adverse events occurring between the first SMQ and the intervention date. This requirement relates to the goal of contesting decisions affecting operations (CL 3.1.3) and the benefit of enhanced ability to identify and address issues (CL 3.2.3).

3. *Change in intervention decision*: Clinicians might require insights into the concordance between the initial management decision made in the SMQ and the final treatment received by the patient. This requirement aligns with the goal of ensuring fair and unbiased AI-driven decision-making (CL 3.1.4) and the benefit of reduced risk of unintended consequences (CL 3.2.4).

These requirements align with the need for explainable techniques that can provide insights into process durations, adverse events, and decision points [53, 482]. Nevertheless, the PM experts should also consider the potential costs and risks for the hospital stakeholders, such as the effort required to interpret and act upon XAI insights (CL 3.3.1) and the challenges in accessing and understanding explanations, particularly for non-expert users (CL 3.3.3) since the hospital team hold different expertise. In this spirit, they should be mindful of possible negative impacts, like over-reliance on AI explanations without human judgment (CL 3.4.1) and information overload (CL 3.4.3).

### 7.2.2 Step 1: Context Mapping

With the explainability requirements established, the next step would be to map these needs to specific subprocesses and decision points within the heart valve disease surgical procedure. The PM experts would work closely with the clinical team to develop detailed process models that capture the key activities, actors, and decision points involved in the process. Through this mapping exercise, the team could identify several critical junctures where explainability would be particularly important:

1. *The pre-operative assessment phase*, where clinicians evaluate the patient's surgical risk based on factors like age, comorbidities, and functional status [271]. Trace attributes (XTA 5.1), event statistics (XTA 5.3), and attribute distributions (XTA 5.4) would be particularly relevant explanation targets in this context.

2. *The selection and sequencing of diagnostic tests*, such as echocardiograms, coronary angiographies, and CT scans, which can impact the overall process duration and resource utilization [390]. Model semantics (XTA 6.2), model metrics (XTA 6.3), and performance analytics (XTA 7.5) would be important explanation targets to consider.

3. *The divergent pathways for elective and emergency cases*, which may involve different test utilization patterns, waiting times, and outcomes [7]. Model variants (XTA 6.4) and process objectives (XTA 7.1) would be relevant explanation targets in this context.

The PM experts should also consider the potential costs and risks for the hospital stakeholders, such as the potential disruption to existing processes and workflows (CL 3.3.2) and increased costs associated with implementing XAI-based process changes (CL 3.3.4). Additionally, they should be aware of possible negative impacts, like resistance to change or adoption of XAI-driven improvements (CL 3.4.2) and lack of informative explanations (CL 3.4.4).

### 7.2.3 Step 2: Need Deconstruction

Based on the elicited requirements, context mapping, and the available data sources, the PM experts would collaborate with the hospital to identify several key explanation targets (6.1.2) spanning the interconnected elements of the PM ecosystem:

- **Event Log Data:** Key targets would include elucidating the 700 variables and 27 declarative constraints, which capture critical information about patient characteristics, clinical events, and process behaviors. Trace attributes

(XTA 5.1), event statistics (XTA 5.3), and attribute distributions (XTA 5.4) would be particularly relevant explanation targets in this context.

- **Process Models:** Essential model-centric targets would include explaining the control-flow and temporal dependencies between the key activities and decision points identified in the context mapping, such as the pre-operative assessment, diagnostic testing, and treatment selection phases. Model semantics (XTA 6.2), model metrics (XTA 6.3), and model variants (XTA 6.4) would be important explanation targets to consider. Another potential area of interest is analyzing the concordance between the initial SMQ decision and the final treatment received by the patient, as well as the reasons for any discrepancies (e.g., changes in patient status or preferences). This aligns with the goal of ensuring fair and unbiased AI-driven decision-making (CL 3.1.4) and the potential benefit of reduced risk of unintended consequences (CL 3.2.4).

- **Business Context:** High-level targets would involve explaining the alignment of the surgical process with the hospital's organizational objectives and constraints, such as patient safety, quality of care, and compliance with internal and external regulations. Process objectives (XTA 7.1), process stakeholders (XTA 7.2), and performance analytics (XTA 7.5) would be relevant explanation targets in this context.

Nevertheless, the PM experts should also consider the potential costs and risks for the hospital stakeholders, such as the effort required to interpret and act upon XAI insights (CL 3.3.1) and increased costs associated with implementing XAI-based process changes (CL 3.3.4). AOnce again, they should be mindful of possible negative impacts, like over-reliance on AI explanations without human judgment (CL 3.4.1) and lack of informative explanations (CL 3.4.4).

### 7.2.4 Step 3: Type Association

Based on the identified decision points and user needs, the PM experts (PD ) would associate specific explanation types to each PM insight, drawing from the taxonomy of explanation types seen in Chapter 4. By focusing on the goals of understanding the AI impact on processes (CL 3.1.2) and contesting decisions affecting operations (CL 3.1.3), the PM experts can ensure that the type association aligns with the needs of the hospital stakeholders in the External User Interaction context.

1. *Descriptive explanations*: Decision trees or rule lists could be used to provide interpretable insights into the patient factors influencing surgical fitness. Data-focused (ETY 8.3), model-focused (ETY 8.4), and rule-based (ETY 9.5) explanations would be particularly relevant in this context.

2. *Contrastive explanations*: Process variant analysis could be applied to highlight the differences between elective and emergency cases, as well as to identify deviations from the ideal process flow. User-based (ETY 9.2) and temporal (ETY 9.3) explanations over the SLA would be important to consider.

3. *Counterfactual explanations*: Techniques like DiCE4EL [206] could be adapted to generate counterfactual scenarios that illustrate how changes in patient characteristics or process steps could lead to different outcomes. Informative (ETY 10.1), cautionary (ETY 10.2), and actionable (ETY 10.5) explanations would be relevant in this context.

### 7.2.5 Step 4: Technical Tool Development

With these explanation types established, the PM experts (PD ) would develop a suite of explainable PM tools tailored to the specific needs of the heart valve disease surgical procedure. By focusing on the goals of understanding the AI impact on processes (CL 3.1.2) and ensuring fair and unbiased AI-driven decision-making (CL 3.1.4), the PM experts can ensure that the technical tool development aligns with the needs of the hospital stakeholders in the External User Interaction context.

For the descriptive explanations, the team could use techniques like one-hot encoding and feature selection to transform the raw process data, including the 700 variables and 27 declarative constraints mentioned in your annotations, into a suitable format for training decision trees or rule-based models. These techniques would be particularly relevant for data explanations (XTA 5, XTA 5.1, XTA 5.3, XTA 5.4) and model explanations (XTA 6, XTA 6.2, XTA 6.3) outlined in Section 6.1.2 of Chapter 6 [467].

For the contrastive explanations, the team could apply trace clustering algorithms (ETY 8.3, ETY 8.4) to group similar patient trajectories based on attributes like demographics, comorbidities, and process steps. These techniques would be particularly relevant for model explanations (XTA 6.4) and business context (XTA 7.1, XTA 7.2) from Section 6.1.2 of Chapter 6 [339].

For the counterfactual explanations, the team could adapt techniques like DiCE4EL [206] to generate plausible scenarios that satisfy the process constraints while mini-

mizing the changes required to achieve a desired outcome. These techniques would align well with the discussion on counterfactual explanations (ETY 9.1) and actionable explanations (ETY 10.5) in Section 6.1.2 of Chapter 6.

However, the PM experts should also consider the potential costs and risks for the hospital stakeholders, such as the effort required to interpret and act upon XAI insights (CL 3.3.1) and increased costs associated with implementing XAI-based process changes (CL 3.3.4). Additionally, they should be mindful of possible negative impacts, like over-reliance on AI explanations without human judgment (CL 3.4.1) and lack of informative explanations (CL 3.4.4).

### 7.2.6 Step 5: Organizational Integration

To ensure that the explainable PM tools are effectively integrated into the hospital's existing workflows and decision-making processes, the PM experts (PD ) would collaborate closely with the hospital's clinical leadership and IT team (CL ) to develop a comprehensive integration plan [39, 204, 445]. By focusing on the goals of understanding the AI impact on processes (CL 3.1.2) and maintaining control over sensitive business data and processes (CL 3.1.5), the PM experts can ensure that the organizational integration aligns with the needs of the hospital stakeholders in the External User Interaction context. The integration process would involve several key steps to ensure seamless data transfer and real-time updates:

1. Data mapping: The team would work with the hospital's IT department to map relevant data fields from the electronic health record (EHR) system to the PM tools. This step would be crucial in addressing potential costs and risks for the hospital stakeholders.

2. API development: To facilitate real-time data transfer between the EHR and the PM tools, the team would develop a set of application programming interfaces (APIs) that allow for secure and automated data exchange [137].

3. User interface integration: The team would collaborate with the hospital's clinical informatics team to embed the explainable PM insights directly into the EHR user interfaces and clinical decision support screens. This step would be essential in tailoring explanations to user mental models and domain knowledge [297, 404], and aligning with the goal of understanding the AI impact on processes (CL 3.1.2).

4. Governance and security: To ensure compliance with relevant regulations, such as the EU General Data Protection Regulation (GDPR) [132] and the Spanish Personal Data Protection and Digital Rights Act (LOPDGDD) [15], the team would work with the hospital's data governance committee to establish clear policies and procedures for data access, use, and retention. This would include implementing strict access controls and data minimization measures to ensure that only authorized personnel can view and use patient data for legitimate purposes, as well as conducting regular data protection impact assessments (DPIAs) to identify and mitigate potential risks to patient privacy. Another key regulation considered in this case for the hospital responsibles would be the EU AI Act. They responsibles should align with the regulatory analysis in Chapter 4, allowing to quantifying better the costs associated with implementing XAI-based process changes (CL 3.3.4) while also identifying the potential risks and impacts related to non-compliance, such as financial penalties (PD 1.4.3) and increased liability (PD 1.4.5) from Section 6.2.2 of Chapter 6.

# Chapter 8

# Conclusion

*Videmus enim nunc per speculum in ænigmate, tunc autem facie ad faciem; nunc cognosco ex parte, tunc autem cognoscam, sicut et cognitus sum.*
— Sancti Pauli Apostoli, *Ad Corinthios, Epistula I 13, 12*

A comprehensive investigation was conducted over explainable AI (XAI) techniques in the context of process mining (PM), aiming to address the research question stated in Section 1.2: *How can XAI techniques and participatory design processes be effectively integrated into PM systems to improve understanding, foster adoption, and ensure solutions that are accurate, usable, compliant, and ethically aligned?*

The systematic literature review in Chapter 2 identified a predominance of feature attribution methods in current XAI approaches for PM, revealing a need for more diverse explanation techniques that can provide causal and contrastive insights. The review also highlighted the scarcity of user-centric evaluations and the limited development of XAI methods tailored to the unique characteristics of PM data and algorithms. Building on these findings, Chapter 3 explored the strategies and barriers faced by PM practitioners in explaining process analyses to clients, uncovering additional gaps related to the alignment of explanations with stakeholder needs, data integration challenges, and the impact of organizational factors on the effectiveness and adoption of explainable PM solutions.

To address these gaps and provide a more comprehensive answer to the research question, the subsequent chapters examined the problem from multiple perspectives.

Chapter 4 investigated the regulatory landscape, emphasizing the need for context-specific explainability approaches that balance transparency and confidentiality requirements. Chapter 5 discussed ethical implications and proposed a risk taxonomy to guide the responsible development and deployment of XAI in PM.

Synthesizing the insights from these chapters, Chapter 6 proposed a conceptual framework that integrates technical, practical, legal, and ethical considerations to operationalize explainability in PM. The framework addresses the identified gaps by advocating for a multi-faceted approach to explanation generation, emphasizing participatory design and iterative refinement based on stakeholder feedback, providing guidance on addressing data integration challenges and organizational constraints, and incorporating legal and ethical considerations into the design and deployment of explainable PM solutions. The hypothetical healthcare case study in Chapter 7 illustrated the potential application of the framework, showcasing how the approach can lead to the development of explainable PM solutions that improve decision-making, patient outcomes, and fairness while navigating policy and ethical challenges.

## Future Work

Building on the insights from this thesis, future research will focus specifically on operationalizing AI governance in business process contexts. This research direction aims to develop practical tools that bridge the gap between high-level ethical principles and policy requirements, alongside their concrete implementation in PM systems.

Key areas of investigation include creating detailed guidelines for conducting ethical impact assessments of PM applications, designing auditable documentation processes for XAI methods in business applications, and developing metrics to quantify compliance with regulations like the EU AI Act. Of particular interest is exploring how to effectively integrate explainability requirements into existing PM methodologies and software tools, ensuring that ethical and legal considerations are addressed throughout the development lifecycle.

In this spirit, future case studies with organizations implementing AI solutions will be conducted to understand the real-world challenges of aligning XAI techniques with governance requirements, and to iteratively refine best practices for responsible AI adoption. This research direction builds upon experience in AI standardization efforts with CEN/CENELEC and practical implementation of ethical AI principles in industry settings.

Nevertheless, it is important to note that the field of XAI in PM is vast and

multifaceted, offering numerous promising avenues for future research. Based on the findings and limitations identified in this thesis, several key directions emerge for the broader research community to explore:

- *Multi-Perspective and Context-Aware Explanations*: The SLR (Chapter 2) revealed a predominance of control-flow focused explanations, highlighting the need for more holistic, multi-perspective approaches that consider data, resources, and context [334, 428]. Future research could explore the integration of additional data sources, such as organizational models [62], decision rules [271], and contextual factors [164], to enrich explanations and provide a more comprehensive understanding of process behavior. Techniques for automatically extracting and aligning multiple perspectives from event logs and domain knowledge could be developed [142, 191], enabling context-aware explanations tailored to specific roles and decision-making scenarios [150, 163].

- *Causal Inference and Counterfactual Explanations in PM*: The SLR also highlighted the need for more diverse explanation techniques beyond feature attribution, particularly causal and contrastive methods. Future research could focus on adapting and developing causal inference algorithms specifically for process mining data and tasks. This could involve extending existing methods like Granger causality [8], structural causal models [403], or reinforcement learning [339] to handle the complex temporal dependencies and multi-dimensional attributes in event logs. Counterfactual explanation methods like DiCE [206] could be tailored to generate realistic, feasible process variants that provide actionable insights for optimization [66, 208]. Additionally, research could explore the integration of causal discovery techniques to support root cause analysis in PM [137, 340].

- *Natural Language Explanations and Conversational Interfaces*: The practitioner study (Chapter 3) emphasized the importance of aligning explanations with stakeholder mental models and enabling interactive querying. Future work could explore the integration of natural language techniques to automatically generate human-readable textual explanations from process mining outputs [374, 481]. This could build upon existing work on Process-to-Text [150] and rule verbalization [190] by leveraging advanced large language models (as strongly remarked by [483]) and incorporating domain-specific terminology [448, 498]. Dialogue systems and conversational agents could be developed

to facilitate exploration of PM insights through intuitive Q&A interactions, drawing inspiration from recent advances in explainable AI chatbots [96] and their applications in PM [102, 187, 506]. User studies evaluating the effectiveness and usability of natural language explanations in PM [26, 163] would provide valuable insights for designing stakeholder-centric solutions.

- *Privacy-Preserving and Secure Explanations*: The regulatory analysis (Chapter 4) surfaced tensions between transparency and confidentiality in high-stakes PM applications. Future research could investigate techniques for generating explanations that protect sensitive information while still providing meaningful insights. This could involve adapting methods from the privacy-preserving ML literature, such as differential privacy [286], federated learning [394], and homomorphic encryption [234] to PM settings. Secure multi-party computation protocols could enable privacy-preserving inter-organizational PM without sharing raw event data [393, 395]. Quantifying the tradeoffs between explanation quality and privacy using novel metrics (e.g., as in [64]) and conducting user studies on the perceived trustworthiness and fairness of privacy-preserving explanations [143, 422] would be important for developing solutions that balance transparency and confidentiality requirements.

- *Simulation and Optimization for Explanation Validation*: Our proposed conceptual framework (Chapter 6, but also before our findings from Chapter 3), highlighted the need for iterative refinement of explanations based on stakeholder feedback. Future work could harness simulation and optimization techniques to efficiently explore the large space of possible process redesigns suggested by explanations [374, 404]. This could involve extending existing PM simulation approaches [252] to incorporate explainability metrics (e.g., as done [444] especially to assess compliance with the EU AI Act) and user preferences as objectives [169, 491]. For example, [339] demonstrated the use of interpretable causal models to optimize process KPIs. Interactive visualization tools tailored to specific business solutions (e.g., [300]) could allow users to navigate and compare alternative process designs, building upon well-defined information needs of process analysts [247]. Case studies evaluating the effectiveness of simulation-based explanation validation in real-world PM projects [143, 449] would provide valuable insights for refining the framework.

- *Ethical AI Governance Frameworks for PM*: The healthcare case study (Chapter 7) illustrated the criticality of proactively addressing explainability needs in

high-stakes PM applications. Departing from our framework in Chapter 6, future research could expand over the proposed ethical principles and risk mitigation strategies (Chapter 5) to match additional governance frameworks and tools for PM [38, 212, 256]. This could involve developing standardized ethical impact assessment instruments tailored to common PM use cases and algorithms [448, 449], as well as auditing and accountability mechanisms, such as data provenance tracking and model factsheets, to ensure adherence to ethical standards [397]. Case studies documenting the challenges and best practices for institutionalizing ethical PM governance, building upon existing industry practices [165, 305, 399], would provide valuable guidance. Integrating ethical considerations into PM methodologies and tools, as advocated by [313, 422], is crucial for responsible AI adoption in process-oriented organizations.

In conclusion, these future research directions span the key themes of the thesis - technical depth, human-centeredness, regulatory compliance, and ethical alignment. This thesis contributes to answering the main research question by identifying critical gaps in current XAI approaches for PM, providing a comprehensive framework that addresses these gaps by integrating multi-disciplinary considerations and offering practical guidance for operationalizing explainability in PM, and demonstrating the potential impact and challenges of applying the framework in a real-world context. By synthesizing insights from multiple perspectives and proposing a holistic approach to explainable PM, this thesis lays the foundation for developing XAI solutions not only accurate but also usable, compliant, and ethically aligned, thereby fostering improved understanding, adoption, and decision-making in PM practice.

# APPENDIX A

# ANALYSES OF EXPLAINABILITY IN AI POLICIES FOR CHAPTER 4

## A.1 Analysis Of International AI Policies for Chapter 4

To ensure a rigorous and comprehensive understanding of the regulatory landscape surrounding AI explainability, we employ a multi-pronged methodological approach that combines thematic analysis with gap analysis. We source policy documents from official governmental and affiliated agencies' websites within the jurisdictions of the European Union (EU), United States (US), and United Kingdom (UK)[1]. We categorize the collected documents into four types:

- Communications: Public statements and releases that outline AI governance strategies.
- Reports: Comprehensive studies, surveys, or official research papers offering in-depth insights.
- Regulations: Legally binding rules and guidelines governing organizational behavior.
- Standards: Technical specifications that detail the implementation guidelines for AI explainability policies.

**Inclusion and Exclusion Criteria** – The documents are selected based on their relevance to the study and the availability of data. We focus on documents produced

[1] For the EU, we refer to official platforms such as the Digital Strategy of the European Commission, Eur Lex, and standard bodies like CEN, CENELEC, and ETSI. For the UK, we consult the official government website and affiliated bodies like ICO and The Alan Turing Institute. For the US, we refer to the White House official website, governmental commissions like NSCAI, and non-regulatory agencies like NIST.

from 2018 onwards to ensure contemporaneity. We exclude documents where the terms "explainability" or "interpretability" are used in contexts not directly related to AI systems, such as auditing procedures.

**Thematic Analysis** – Our thematic analysis aims to identify common themes across the collected documents. We focus on key concepts like "explainability," "transparency," and "trustworthiness," and report on the similarities and differences in how these concepts are accounted for across jurisdictions. This analysis is guided by a coding scheme that categorizes the data into thematic clusters.

**Gap Analysis** – To identify gaps in the current regulatory landscape, we compare the themes derived from policy documents with existing academic literature from various research communities, including algorithmic studies, human-computer interaction, and ethics. This allows us to pinpoint areas where policies may be misaligned with current research, thereby identifying avenues for future work.

### A.1.1 Tables of Policy References Consulted in Chapter 4

In the following page, Table 1 reports a non-exhaustive list of policy documents such as communications, reports, and regulations involving AI Explainability. For each document is reported year of announcement, while within the Regulations columns, square brackets denote status.

Table 2 reports a comprehensive representation of major standardization activities affecting AI Explainability. The asterisk sign denotes either expected publication release or, if delayed, latest announcement for expected publication release. Note that *INT* in the Area column refers to international standardization organizations, while the column *Delegation* refers to specific committee and/or working groups responsible for standard documents provision.

**Table A.1:** Policy communications, reports, and regulations affecting AI explainability.

| Area | Communications | Reports | Regulations |
|---|---|---|---|
| **EU** | • 2018, EC AI for Europe [85]<br>• 2020, White Paper on AI [86]<br>• 2021, EC Fostering a European approach to AI [87]<br>• 2021, EC Coordinated Plan on AI 2021 Review [88]<br>• 2022, EC Rolling Plan for ICT Standardisation - AI [89] | • 2019, HLEG Guidelines for Trustworthy AI [369]<br>• 2020, HLEG Assessment List for Trustworthy Artificial Intelligence (ALTAI) [370]<br>• 2021, JRC AI Standardization Landscape [90] | • 2018, EU GDPR [Enactment] [132]<br>• 2021, AI Act [First Draft] [358]<br>• 2022, AI Liability Directive [Proposal] [359]<br>• 2022, AI Act [Final Draft - General Approach] [360] |
| **US** | • 2019, WH - Executive Order on Maintaining American Leadership in AI [361]<br>• 2020, WH - OMB - Guidance for Regulation of AI Applications [362]<br>• 2022, Blueprint for AI Bill of Rights [355] | • 2021, NSCAI Final Report [371]<br>• 2021, NIST Federal Engagement Plan [356]<br>• 2021, NISTIR 8367 [60]<br>• 2021, NISTIR 8312 [387]<br>• 2023, NIST AI RMF v1.0 [357] | • 2021, National AI Initiative Act [Enactment] [2]<br>• 2022, Algorithmic Accountability Act [Draft] [3] |
| **UK** | • 2019, National Data Strategy [152]<br>• 2019, AI Sector Deal [153]<br>• 2019, ICO & Alan Turing Institute Project ExplAIn [215]<br>• 2022, National AI Strategy - AI Action Plan [155]<br>• 2022, Establishing a pro-innovation approach to regulating AI [154] | • 2020, ICO, Guidance on the AI Auditing Framework [364]<br>• 2020, ICO, Guidance on AI and Data Protection [363]<br>• 2020, ICO & Alan Turing Institute, Explaining Decisions with AI [216]<br>• 2022, Alan Turing Institute, Common Regulatory Capacity for AI [18] | • 2018, Data Protection Act [Enactment] [365] |

**Table A.2:** Standards affecting AI explainability.

| Area | Standardization Body | Delegation | Document(s) | Publication | Release Date |
|---|---|---|---|---|---|
| EU | **CEN - CENELEC** | JTC 21 AI [75] | Explainability, Verifiability [74] | Proposal | TBD |
| | **ETSI** | ISG SAI [125] | DGR/SAI-007 [136, 335] | Under Appr. | 2023-03* |
| | | | DGR/SAI-010 [335] | Proposal | TBD |
| | | GS ENI | GS ENI 005 v2.1.1 [135, 335] | Published | 2021 |
| US | **NIST** | NIST Interagency | NISTIR 8312 (Paper) [387] | Published | 2021-09 |
| | | | NISTIR 8367 (Paper) [60] | Published | 2021-04 |
| | | | RMF V1.0 [357] | Published | 2023-01 |
| UK | **BSI - NPL** | AI Standard Hub (CDDO, CDEI) | Algorithmic Transparency Recording Standard (Guide) [76] | Published | 2021-12 |
| | | CDDO, Cabinet Office, Office for AI | Ethics, Transparency and Acc. Framework for ADM (Guide) [367] | Published | 2021-05 |
| INT | **ISO/IEC** | JTC 1/SC 42 AI [217] | ISO/IEC AWI 12792 [222] | Under Dev. | 2025-02 |
| | | | ISO/IEC AWI TS 6254 [218] | Under Dev. | 2024-02 |
| | | | ISO/IEC TR 24028:2020 [221] | Published | 2020-05 |
| | **IEEE** | CIS/SC/XAI WG | P2976 [156] | Inititation | 2024-07* |
| | | VT/ITS | 7001-2021 [213] | Published | 2022-03 |
| | | C/AISC/XAI | P2894 (Guide) [36] | Under Dev. | 2022-03* |
| | | C/S2ESC/ALGB-WG | P7003 [439] | Published | 2017-02 |

## A.2 Coding Analysis on Explainability in EU AI Policies for Chapter 4

This section provides supplementary information detailing the methodology used for the policy analysis presented in the manuscript "Operationalizing Explainable AI in the EU Regulatory Ecosystem" that informed Chapter 4. The purpose is to enhance transparency and rigor by thoroughly describing our qualitative approach to identifying, extracting, and analyzing legal explainability requirements across European Union (EU) regulations related to data, artificial intelligence (AI) systems, and digital platforms. The methodology is structured along the following stages:

- *Policy Selection*: Database used and process for identifying relevant EU regulations to analyze
- *Data Extraction*: How relevant articles and recitals were identified from the legal texts
- *Data Analysis*: Inductive process of reviewing extracts to map explainability provisions to dimensions and stakeholders
- *Discussion of Limitations*: Notes constraints and subjectivity inherent to qualitative legal analysis
- *Legislative Sections*: For each legal text as subsection under the *Coding* section, there are subsubsections containing analysis of provisions alongside a table reporting their verbatim.

By comprehensively outlining each step of our methodology, we aim to provide full transparency into how we systematically identified legal explainability mandates situated across the EU policy ecosystem.

### A.2.1 Policy Selection

Our policy analysis focuses exclusively on regulations and legislative proposals from official European Union (EU) institutions, particularly the European Commission, Parliament, and Council. The EurLex database was used as the sole repository to retrieve policies. EurLex contains the official and definitive versions of all EU legal texts, including treaties, legislation, case law, and legislative proposals. Using EurLex ensured we analyzed the most current and authoritative regulatory texts. We conducted

a structured search of *EurLex* to identify relevant regulations and proposals within the domains of data, artificial intelligence (AI), and digital platforms. These three areas represent the core topics shaping explainability requirements for algorithmic systems under the EU's digital policy agenda.
The search utilized variations of the following terms: *"artificial intelligence", "AI", "algorithm", "automated decision", "digital", "platform", "data", "transparency",* and *"explainability"*. Filters were applied to narrow results to binding legislation and impactful proposals from EU institutions. Document Type was limited to "Regulation" or "Directive" or proposals of such typologies; this excluded other types like Decisions, Legislative Resolutions, Cover notes, Provisional Data etc. Results were screened for inclusion based on containing provisions explicitly addressing explainability, interpretability, transparency or related concepts like traceability and auditability for AI/algorithmic systems. Texts focused narrowly on non-algorithmic manners with no clear connection to AI explainability were excluded after a first read-through screening. As an example, considered but excluded regulations were found such as the *Regulation on Markets in Crypto-assets (2022/2555)*; *Regulation on semiconductor ecosystem (2023/1114)*, and the *Directive on cybersecurity measures (2023/1781)*. In detail, the search results were overall screened via a two-step process:

(**1.**) First, title/abstract screening was conducted to identify potentially relevant regulations and proposals containing provisions related to explainability, interpretability or transparency of algorithmic systems. At this stage, several texts were identified, including the General Data Protection Regulation (GDPR), Artificial Intelligence Act (AIA), AI Liability Directive, Digital Services Act (DSA), Digital Markets Act (DMA), Data Governance Act and Data Act and the aforementioned regulations and directives on crypto-assets, semiconductors, and cybersecurity.

(**2.**) Second, the full texts of these initial candidates were reviewed to determine if they contained explicit legal requirements for explainability of AI or algorithmic systems relevant to our analysis aims. During this second screening, the Data Governance Act and Data Act were excluded, as further examination revealed they did not directly establish explainability mandates for AI systems. While important for data governance, provisions were deemed insufficiently related to transparency of automated decision-making systems.

Despite initially identifying other candidates, careful screening ensured only texts centrally shaping legal explainability requirements were analyzed. The final policies included for analysis are:

- *General Data Protection Regulation* (GDPR) – 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) (Text with EEA relevance)

- *Artificial Intelligence Act* (AIA) – Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 2 February 2024

- *AI Liability Directive proposal* – Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)

- *Digital Services Act* (DSA) – Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)

- *Digital Markets Act* (DMA) – Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)

### A.2.2 Data Extraction

This set comprehensively covers the EU's regulatory foundation shaping algorithmic explainability requirements and tensions. The seminal nature and definitive status of these texts within EU digital policy justified their inclusion. The full legal texts of the selected EU regulations and proposals were retrieved from EurLex in PDF format. For consolidated acts like the GDPR, the most recent amended version was used.

A qualitative content analysis approach was utilized to identify and extract provisions related to explainability requirements. The PDF policy documents were manually reviewed to pinpoint articles and accompanying recitals containing explicit mandates or stipulations regarding interpretability, transparency, explainability or related concepts like traceability and auditability for AI systems.

No qualitative data analysis software was used - relevant sections were highlighted directly within the PDF documents and tagged based on the specific legal explainability requirement. An iterative process was conducted by a PhD student and AI governance analyst in XAI. The process sought to develop preliminary codes and categories for organizing the extracts into broader explainability themes and dimensions. For example,

provisions were broadly coded as pertaining to algorithmic data practices, AI system processes, or organizational/business practices requiring transparency. Additional researchers [full professors in XAI, AI, HCI, and business process management] were included to discuss and validate a reached consensus over explainability dimensions and stakeholders.

### A.2.3 Data Analysis

A primarily inductive approach was used to analyze the extracted explainability requirements without applying a formal grounded theory or external coding framework. The tagged legal provisions were iteratively reviewed to identify common themes, dimensions, and tensions. Through this process, the following codes for explainability dimensions and stakeholders emerged:

- Explainability Dimensions:
    1. **Data** - provisions related to data practices, processing, and outputs
    2. **Process** - provisions on AI system architecture, processes, and decision-making
    3. **Business** - provisions on organizational practices, design choices, and deployment
- Stakeholders:
    1. **Provider** - entities like developers and deployers responsible for AI systems
    2. **User** - individuals affected by AI system decisions; data subjects and claimants
    3. **Auditor** - oversight bodies and legal entities conducting compliance audits

Additionally, stakeholders were normally mapped conceiving a '*giver*' and a '*recipient*' of explanations; in some cases, yet, we defined '*interpreter*' whenever the articles were not redirecting to any specific explainer. These codes capture the core aspects of explainability covered in the EU regulations. The dimensions reflect the different targets of explainability requirements, from data handling to system processes and business practices. The stakeholder codes encompass the key entities involved in generating or receiving explanations across the AI system lifecycle.

Key themes also emerged from grouping related requirements, including:

1. **Oversight Procedures** - clustering provisions pertaining to auditing and monitoring AI systems for compliance
2. **End-User Services** - gathering requirements focused on providing transparency and explanations to users affected by AI decisions

3. **Legal Liability** Scenarios - identifying provisions related to explainability in the context of litigation and establishing liability for AI harms

These codes and categories evolved through close re-reading of the legal extracts to capture common explainability mandates and tensions within the EU policy ecosystem. They provided a framework for systematically mapping and synthesizing the explainability requirements situated across the various regulations.

The mapping was summarized in Table 1, pag.7 of the manuscript. It highlights relevant articles, recitals, dimensions and stakeholders addressed in each EU regulation. This table synthesis aimed to provide a clear overview of legal explainability mandates situated across policies. To promote consistent interpretation and bolster validity, the mapping was discussed among the team and underwent multiple iterations based on independent review. Coding discrepancies were resolved through re-examination and deliberation to reach consensus. While fundamentally an inductive and subjective process, seeking agreement on categorization from multiple analysts enhanced rigor by minimizing individual bias in analyzing the legal explainability requirements.

The following sections represent each a legislative text, where the detected articles and recitals are paired if considered relevant in their mentions. As further addition, *Explainability Dimension(s)* and *Stakeholder(s)* involved report the outcomes of the Coding Iteration column.

### A.2.4 Discussion of Limitations

While we aimed to systematically identify and map legal explainability requirements in an objective manner, some inherent limitations should be acknowledged:

- Subjectivity in interpreting legal texts - As a qualitative analysis, categorizing explainability provisions involves inherent subjectivity. Terms like “meaningful information" are ambiguous and open to interpretation. We sought to minimize bias by resolving discrepancies through discussions and consensus. However, absolute objectivity cannot be guaranteed.

- Scope limited to specific regulations - Our scope was narrowed to key EU policies on data, AI and platforms. Additional national laws and sector-specific regulations may contain further explainability mandates not examined.

- Legal certainty - As academic researchers in XAI and AI policy, our analysis provides an informed perspective on the legal explainability requirements.

> However, the definitive interpretation and application of these EU regulations ultimately lies with the Court of Justice of the European Union and other relevant legislative and judicial bodies. Our analysis should be understood as an academic viewpoint to map the explainability landscape, rather than authoritative legal judgements.

We now present the detailed coding process used to analyze the legal explainability requirements in each EU regulation. The full text of relevant articles and recitals was examined, and initial codes were proposed for the explainability dimensions and stakeholders addressed in each provision. These codes were then discussed, iteratively refined, and consensus was reached on the final categorizations.

### A.2.5 General Data Protection Regulation – 32016R0679 - EN - EUR-Lex

#### Articles 13(2)(f),14(2)(g),15(1)(h)

Upon initial examination of these articles, the following codes were proposed:

- Dimension: Data. The focus seems to be on providing information about the personal data processing and profiling logic to the data subject.
- Stakeholder: User, Provider. The information is to be given by the provider/-controller to the data subject (user).

Further analysis yielded additional codes:

- Dimension: Data, Process. The articles mention both personal data and the automated decision system logic.
- Stakeholder: User. The information goes to the data subject/user.

Upon discussion, it was agreed that Process should also be included as a dimension since the articles explicitly refer to the logic of the automated systems. It was also agreed that Provider should be a designated stakeholder given they are responsible for supplying the information about the systems and data processing.

The final consensus codes were:

- Dimension: Data, Process, Business. Covers both personal data and automated system logic.

- Stakeholder: User, Provider

Providers supply information to data subject users.

The related Recital 60 further confirms the focus on informing data subjects about profiling and consequences of data processing (therefore likely business impacts) in a transparent manner.

| **Art.13(2)(f),14(2)(g),15(1)(h)** | **Recital 60** |
|---|---|
| *Article 13 Information to be provided where personal data are collected from the data subject 2. In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing:* ***(f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject;***<br>*Article 14 Information to be provided where personal data have not been obtained from the data subject 2. In addition to the information referred to in paragraph 1, the controller shall provide the data subject with the following information necessary to ensure fair and transparent processing in respect of the data subject:* ***(g) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.***<br>*Article 15 Right of access by the data subject 1. The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and the following information:* ***(h) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.*** | *The principles of fair and transparent processing require that the data subject be informed of the existence of the processing operation and its purposes. The controller should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed. Furthermore,* ***the data subject should be informed of the existence of profiling and the consequences of such profiling. Where the personal data are collected from the data subject, the data subject should also be informed whether he or she is obliged to provide the personal data and of the consequences, where he or she does not provide such data. That information may be provided in combination with standardised icons in order to give in an easily visible, intelligible and clearly legible manner, a meaningful overview of the intended processing.*** *Where the icons are presented electronically, they should be machine-readable.* |

### Article 22(3)

- Dimension: Process. The focus of the article is on the automated decision-making system and profiling processes, rather than personal data itself.
- Stakeholder: User, Provider
  - The data subject (user) has rights related to the system's outputs that affect them.

– The provider operates the automated processing system.

Further analysis of Article 22(3) yielded:

- Dimension: Data, Process. The automated system is processing user's personal data, so data is a factor. However, the core focus is the automated decision process.
- Stakeholder: User. The rights pertain to the user subjected to the automated decision output.

On discussion, it was agreed that Process better captures the core focus on the automated systems, rather than data. It was also conceded that Providers are implicitly involved as the operators of such systems, even if not explicitly mentioned.

The final consensus codes were:

- Dimension: Process, Data
- Stakeholder: User, Provider

### A.2.6 AI Act – 2021/0106(COD) 5662/24 - EN - EUR-Lex

#### Article 13(1),(3)

Article 13 focuses on ensuring AI system transparency and interpretability foremost for providers.

Initial codes proposed were:

- Dimension: Process. The transparency and explainability requirements focus on the AI system itself, how it functions and makes decisions. (Article 13, Recitals)
- Dimension: Data. The information on system capabilities covers things like accuracy metrics which pertain to input/training data. (Art. 13(3)(b))
- Stakeholder: Provider, Deployer, User
- Transparency facilitates oversight and control over the system by all relevant stakeholders.

Further analysis yielded:

| **Art.22(3)** | **Recital 71** |
|---|---|
| *1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*<br>*2. Paragraph 1 shall not apply if the decision:*<br>1. *is necessary for entering into, or performance of, a contract between the data subject and a data controller;*<br>2. *is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or*<br>3. *is based on the data subject's explicit consent.*<br>*3. In the cases referred to in points (a) and (c) of paragraph 2,* ***the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.*** | *The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent.* ***In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. Such measure should not concern a child.*** |

- Dimension: Process, Business. Transparency applies to the technical system as well as organizational practices around deployment. (Art 13, Recitals)
- Dimension: Data. Accuracy metrics relate to the training data.
- Stakeholder: Provider, Deployer. The legal obligations around transparency lie with providers to enable deployers, focus less on end users.

On discussion, it was agreed that end user oversight is intended as an outcome of transparency requirements even if users don't have direct legal obligations. It was also conceded that business processes would reasonably be impacted by the transparency directives. End users seem not legally enabled to oversee system transparency.

The final consensus codes were:

- Dimension: Process, Data

- Stakeholder: Provider, Deployer, User

### Article 14(4)

Initial codes for Article 14(4) were:

- Dimension: Process. Primary focus is explaining the system's capabilities and functions for oversight (Article 14)
- Stakeholder: User. Enabling understanding for the human operators assigned oversight roles.

Further analysis yielded:

- Dimension: Business. Organizational oversight measures are important explainability requirements. (Recital 48)
- Dimension: Data. System outputs need interpretation too. (Art. 14(4)(c))
- Stakeholder: Provider. Providers remain responsible for designing to support oversight.

On deliberation, it was agreed that organizational aspects are relevant for oversight even if not explicitly stated. It was also conceded that human operators are the main stakeholders discussed in the texts. Yet, the focus is on the model's data output to be interpretable. The concept of 'natural person' seems here to refer to providers or deployers, aka people assigned to monitoring the AI model, thus apparently not focusing on end-user's interpretability per se.

Final Consensus Codes

- Dimension: Data, Process
- Stakeholder: Provider

### Article 50(1)

Based on Article 50 and Recital 70a, two perspectives on explainability goals related to disclosing AI systems to users were found.

Initial codes proposed were:

| | |
|---|---|
| **Article 13(1)** | *High-risk AI systems shall be designed and developed in such a way to ensure that their operation is* ***sufficiently transparent to enable deployers to interpret the system's output and use it appropriately.*** *An appropriate type and degree of transparency shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Chapter 3 of this Title.* |
| **Article 13(2)** | ***High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to users.*** |
| **Article 13(3)(b)(ii)** | ***the characteristics, capabilities and limitations of performance of the high-risk AI system, including: (i) its intended purpose; (ii) the level of accuracy, including its metrics, robustness and cybersecurity*** *referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;* |
| **Article 13(3)(b)(iiia)** | ***where applicable, the technical capabilities and characteristics of the AI system to provide information that is relevant to explain its output****;* |
| **Article 13(3)(d)** | *the human oversight measures referred to in Article 14, including* ***the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the deployers***; |
| **Recital 9b** | *In order to obtain the greatest benefits from AI systems while protecting fundamental rights, health and safety and to enable democratic control, AI literacy should equip providers, deployers and affected persons with the necessary notions to make informed decisions regarding AI systems. These notions may vary with regard to the relevant context and can include understanding the correct application of technical elements during the AI system's development phase, the measures to be applied during its use,* ***the suitable ways in which to interpret the AI system's output, and, in the case of affected persons, the knowledge necessary to understand how decisions taken with the assistance of AI will impact them.*** |
| **Recital 14a** | *While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the 2019 Ethics Guidelines for Trustworthy AI developed by the independent High-Level Expert Group on AI (HLEG) appointed by the Commission. In those Guidelines [...]* ***human agency and oversight means that AI systems are developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans. [...] Transparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights.*** |
| **Recital 47** | ***High-risk AI systems should be designed in a manner to enable deployers to understand how the AI system works, evaluate its functionality, and comprehend its strengths and limitations.*** *High-risk AI systems should be accompanied by appropriate information in the form of instructions of use. Such information should include the characteristics, capabilities and limitations of performance of the AI system [...]* ***Transparency, including the accompanying instructions for use, should assist deployers in the use of the system and support informed decision making by them.*** *Among others, deployers should be in a better position to make the correct choice of the system they intend to use in the light of the obligations applicable to them, be educated about the intended and precluded uses, and use the AI system correctly and as appropriate.* ***In order to enhance legibility and accessibility of the information included in the instructions of use, where appropriate, illustrative examples, for instance on the limitations and on the intended and precluded uses of the AI system, should be included. Providers should ensure that all documentation, including the instructions for use, contains meaningful, comprehensive, accessible and understandable information, taking into account the needs and foreseeable knowledge of the target deployers****.* |

| **Art.14(4)(c)** | **Recital 48** |
|---|---|
| [...] *the high-risk AI system shall be provided to the user in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate to the circumstances: (a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, also in view of detecting and addressing anomalies, dysfunctions and unexpected performance; (b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias'), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;* ***(c) to correctly interpret the high-risk AI system's output, taking into account for example the interpretation tools and methods available;*** | *High-risk AI systems should be designed and developed in such a way that natural persons can oversee their functioning, ensure that they are used as intended and that their impacts are addressed over the system's lifecycle. For this purpose, appropriate human oversight measures should be identified by the provider of the system before its placing on the market or putting into service. In particular,* ***where appropriate, such measures should guarantee that the system is subject to in-built operational constraints that cannot be overridden by the system itself and is responsive to the human operator, and that the natural persons to whom human oversight has been assigned have the necessary competence, training and authority to carry out that role. It is also essential, as appropriate, to ensure that high-risk AI systems include mechanisms to guide and inform a natural person to whom human oversight has been assigned to make informed decisions if, when and how to intervene in order to avoid negative consequences or risks, or stop the system if it does not perform as intended.*** |

- Dimension: Process. Detecting AI interactions and system outputs supports process transparency. (Article 50)
- Stakeholder: User. Important for informing end users when engaging with AI systems.

Further analysis yielded:

- Dimension: Business. Organizational practices around deploying biometric/deepfake AI systems are regulated. (Article 50)
- Dimension: Data. Tracing origin of system outputs requires technical data provenance solutions. (Recital 70a)
- Stakeholder: Provider. Responsibility on providers to enable detection/disclosure of AI systems and outputs.

Upon discussion, it was agreed that business processes are affected by the AI disclosure directives. It was also conceded that end user awareness is still a goal, despite focus on providers. Disclosure supports process transparency and traceability to engender user trust in AI systems.

Final Consensus Code

- Dimension: Process, Business

- Stakeholder: User, Provider

| | |
|---|---|
| **Article 50(1)** | ***Providers shall ensure that AI systems intended to directly interact with natural persons are designed and developed in such a way that the concerned natural persons are informed that they are interacting with an AI system, unless this is obvious from the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account the circumstances and the context of use****. This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties unless those systems are available for the public to report a criminal offence.* |
| **Article 50(3)** | *Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated. This obligation shall not apply where the use is authorised by law to detect, prevent, investigate and prosecute criminal offence.* ***Where the content forms part of an evidently artistic, creative, satirical, fictional analogous work or programme, the transparency obligations set out in this paragraph are limited to disclosure of the existence of such generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work.*** |
| **Recital 70a** | *variety of* ***AI systems can generate large quantities of synthetic content*** *that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem, raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception. In the light of those impacts, the fast technological pace and the need for new methods and techniques to trace origin of information,* ***it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints or other techniques, as may be appropriate.*** |

### Article 86

Initial codes proposed for Article 86 were:

- Dimension: Process. Explanations pertain to how AI system outputs factor into decisions. (Art. 86)

- Stakeholder: User. Important transparency right for affected persons to understand outcomes.

Further analysis yielded:

- Dimension: Business. Deployer practices for decisions based on AI outputs are regulated. (Art. 86)
- Dimension: Data. Dataset factors could be relevant to judging system outputs behind decisions.
- Stakeholder: Provider. Though not stated, providers design systems deployed in such applications.

On deliberation, it was agreed that business processes are reasonably impacted. It was also conceded that end users have an explicit right even if providers are only implicitly covered. The analysis aligned that, based on output data, end-users need to receive an explanation of the model's rationale and envisaged consequences. Explanation rights support oversight over automated decisions, i.e., it is expected to align with, not replace or impede, what is already expressed within the GDPR.

Final Consensus Codes

- Dimension: Process, Business, Data
- Stakeholder: User, Provider

| **Art.86** | **Recital 84b** |
|---|---|
| *1.* ***Any affected person subject to a decision which is taken by the deployer on the basis of the output from an high-risk AI system listed in Annex III, with the exception of systems listed under point 2, and which produces legal effects or similarly significantly affects him or her in a way that they consider to adversely impact their health, safety and fundamental rights shall have the right to request from the deployer clear and meaningful explanations on the role of the AI system in the decision-making procedure and the main elements of the decision taken****. 2. Paragraph 1 shall not apply to the use of AI systems for which exceptions from, or restrictions to, the obligation under paragraph 1 follow from Union or national law in compliance with Union law. 3. This Article shall only apply to the extent that the right referred to in paragraph 1 is not already provided for under Union legislation.* | ***Affected persons should have the right to request an explanation when a decision is taken by the deployer with the output from certain high-risk systems as provided for in this Regulation as the main basis and which produces legal effects or similarly significantly affects him or her in a way that they consider to adversely impact their health, safety or fundamental rights. This explanation should be a clear and meaningful and should provide a basis for affected persons to exercise their rights****. This should not apply to the use of AI systems for which exceptions or restrictions follow from Union or national law and should apply only to the extent this right is not already provided for under Union legislation.* |

### General Purpose AI Models

The approved EU AI Act text refers to general purpose AI models, termed GPAI models (Art. 50), which have the capability to competently perform a wide range of tasks[2]. While not formally classified as high-risk, GPAI models still face notable transparency requirements around responsible disclosures for providers to downstream users.

Specifically, Article 52c mandates GPAI providers maintain technical documentation including acceptable use policies, tasks the model can perform, data used for training/testing, computational resources utilized, and other relevant information. This documentation must be provided upon request to authorities and made available to downstream providers integrating the GPAI model into their own systems to facilitate responsible usage. Elements to disclose align closely with the transparency rules in Article 13 for high-risk systems.

Additionally, GPAI models presenting systemic risks incur extra transparency duties per Article 52d resembling high-risk systems, such as conducting adversarial testing, documenting model evaluations on public protocols, detailing system architecture, and applying cybersecurity protections.

In this vein, Recital 60e further stress 'proportionate transparency measures' for providers of general-purpose models, including maintaining documentation and providing usage information to downstream providers, while Recital 60f defines transparency exceptions for general-purpose models under free and open source license if they do not carry systematic risks. As a support, Annex IXb provides detailed list of information required for primary transparency obligations to downstream providers as set in Art.52c.

However, GPAI models presently do not undergo the formal benchmarking procedures or accuracy validations required of high-risk systems in Article 15. Metrics indicating capabilities must be reported, but concrete thresholds are not outlined.

In conclusion, while specifics differ and remain less extensive compared to comprehensive high-risk systems transparency rules, general-purpose AI models still face significant transparency obligations around disclosures for responsible usage under the EU AI Act regime. But performance expectations appear lighter currently and expan-

[2]The EU AI Act defines general purpose AI models (GPAI) as systems capable of competently performing a wide range of tasks. Article 50 spans both high-risk as well as GPAI systems interacting with humans or generating synthetic audio/video/text content. For GPAI - classified in Art.52a - it details requirements mandating transparency documentation disclosing acceptable uses, data/training processes, model capabilities/limitations, and content used in Art52c. But specifics differ and remain less extensive presently compared to comprehensive rules for high-risk systems.

sion of transparency duties for GPAI models seems likely as technical understanding of systemic risks develops further.

### A.2.7 AI Liability Directive – 52022PC0496 - EN - EUR-Lex

#### Article 3(1)

Article 3 allows courts to order providers to disclose evidence about a high-risk AI system suspected of causing harm, upon request from a claimant.

Initial codes proposed were:

- Dimension: Process. Focuses on evidence about how the AI system functions.
- Stakeholder: Auditor, Provider, User
  - Courts can compel providers to disclose info.
  - Users are claimants requesting info.

Further analysis yielded:

- Dimension: Business. Relates more to organizational practices and liability.
- Stakeholder: Auditor, Provider. Courts and providers are involved.

Through discussion, it was agreed that Business better captures the focus on liability procedures and evidence disclosure practices, rather than just technical explainability, which may yet warrant investigation. It was also conceded that Users should be included as claimants petitioning for information.

The final consensus was:

- Dimension: Business, Process
- Stakeholder: Auditor, Provider, User

#### Article 4(4),(5)

This article discusses applying a presumption of causation between an AI system and harm, based on the level of explainability and available evidence.

Initial codes proposed were:

- Dimension: Process. Explainability of the AI system's functioning is key.

| Art.3(1) | Recital 19 |
|---|---|
| *Disclosure of evidence and rebuttable presumption of non-compliance Member States shall ensure that national courts are empowered, either upon the request of a potential claimant who has previously asked a provider, a person subject to the obligations of a provider pursuant to [Article 24 or Article 28(1) of the AI Act] or a user* ***to disclose relevant evidence at its disposal about a specific high-risk AI system that is suspected of having caused damage, but was refused****, or a claimant, to order the disclosure of such evidence from those persons.* | *In order for the judicial means to be effective, Article 3(3) of the Directive provides that a court may also order the preservation of such evidence. As provided in Article 3(4), first subparagraph,* ***the court may order such disclosure, only to the extent necessary to sustain the claim, given that the information could be critical evidence to the injured person's claim in the case of damage that involve AI systems.*** |

- Stakeholder: Auditor, Provider, User. Evaluation done by courts based on evidence from providers and claimants (users).

Further analysis yielded:

- Dimension: Business. Relates more to liability procedures and evidentiary requirements.
- Stakeholder: Auditor, Provider. Courts and providers are involved.

After discussion, it was agreed that Business should be the primary focus given the articles relate to legal liability practices more than technical explainability itself.

However, Process was still a relevant dimension in considering required evidence about the AI system, per Recital 28.

It was also agreed to include Users as claimants, though their role is more indirect.

The final consensus was:

- Dimension: Process, Business
- Stakeholder: Auditor, Provider, User

While Recitals 22-30 can be deemed as relevant, particularly Recital 28 further highlights opacity hampering claimants in proving AI system causation

### A.2.8 Digital Service Act – 32022R2065 - EN - EUR-Lex

#### Article 27(1),(2)

Article 27 requires online platforms using recommender systems to explain the main parameters used to users.

Initial codes proposed were:

| **Art.4** | **Recital 28** |
|---|---|
| Art.4(4),(5): Rebuttable presumption of a causal link in the case of fault<br>*4. In the case of a claim for damages against a provider of a high-risk AI system subject to the requirements laid down in chapters 2 and 3 of Title III of [the AI Act] or a person subject to the provider's obligations pursuant to [Article 24 or Article 28(1) of the AI Act], the condition of paragraph 1 letter (a) shall be met only where the complainant has demonstrated that the provider or, where relevant, the person subject to the provider's obligations, failed to comply with any of the following requirements laid down in those chapters, taking into account the steps undertaken in and the results of the risk management system pursuant to [Article 9 and Article 16 point (a) of the AI Act]: (a) the AI system is a system which makes use of techniques involving the training of models with data and which was not developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in [Article 10(2) to (4) of the AI Act];* ***(b) the AI system was not designed and developed in a way that meets the transparency requirements laid down in [Article 13 of the AI Act]; (c) the AI system was not designed and developed in a way that allows for an effective oversight by natural persons during the period in which the AI system is in use pursuant to [Article 14 of the AI Act];*** *5. In the case of a claim for damages concerning an AI system that is not a high-risk AI system, the presumption laid down in paragraph 1 shall only apply where the national court considers it excessively difficult for the claimant to prove the causal link mentioned in paragraph 1.* | *The presumption of causality could also apply to AI systems that are not high-risk AI systems because there could be excessive difficulties of proof for the claimant. For example, such difficulties could be assessed in light of the characteristics of certain AI systems, such as autonomy and* ***opacity, which render the explanation of the inner functioning of the AI system very difficult in practice, negatively affecting the ability of the claimant to prove the causal link between the fault of the defendant and the AI output. A national court should apply the presumption where the claimant is in an excessively difficult position to prove causation, since it is required to explain how the AI system was led by the human act or omission that constitutes fault to produce the output or the failure to produce an output which gave rise to the damage.****However, the claimant should neither be required to explain the characteristics of the AI system concerned nor how these characteristics make it harder to establish the causal link.* |

- Dimension: Process. Focuses on how the recommender system functions.
- Stakeholder: Provider, User. Platforms must provide explanations to users.

Further analysis yielded:

- Dimension: Business. Relates more to platform information practices.
- Stakeholder: Provider, User. Platforms explain to users.

Through discussion, it was agreed that Business better captures the transparency requirement for platforms' organizational practices rather than just technical explanations. Secondary explainability targets can be considered user data as elaborated by recommender systems. The final codes were:

- Dimension: Business
- Stakeholder: Provider, User

Recital 68 further emphasizes platforms should provide meaningful explanations of profiling logic used for targeted advertising.

| **Art.27(1)(2)** | **Recital 68** |
|---|---|
| *Recommender system transparency*<br>1. ***Providers of online platforms that use recommender systems shall set out in their terms and conditions, in plain and intelligible language, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters.***<br>2. ***The main parameters referred to in paragraph 1 shall explain why certain information is suggested to the recipient of the service. They shall include, at least:***<br>***the criteria which are most significant in determining the information suggested to the recipient of the service;***<br>***(b) the reasons for the relative importance of those parameters.*** | *Online advertising plays an important role in the online environment, including in relation to the provision of online platforms, where the provision of the service is sometimes in whole or in part remunerated directly or indirectly, through advertising revenues. Online advertising can contribute to significant risks, ranging from advertisements that are themselves illegal content, to contributing to financial incentives for the publication or amplification of illegal or otherwise harmful content and activities online, or the discriminatory presentation of advertisements with an impact on the equal treatment and opportunities of citizens. In addition to the requirements resulting from Article 6 of Directive 2000/31/EC, providers of online platforms should therefore be required to ensure that the recipients of the service have certain individualised information necessary for them to understand when and on whose behalf the advertisement is presented. They should ensure that the information is salient, including through standardised visual or audio marks, clearly identifiable and unambiguous for the average recipient of the service, and should be adapted to the nature of the individual service's online interface.* ***In addition, recipients of the service should have information directly accessible from the online interface where the advertisement is presented, on the main parameters used for determining that a specific advertisement is presented to them, providing meaningful explanations of the logic used to that end, including when this is based on profiling.***<br>***Such explanations should include information on the method used for presenting the advertisement, for example whether it is contextual or other type of advertising, and, where applicable, the main profiling criteria used; it should also inform the recipient about any means available for them to change such criteria****. The requirements of this Regulation on the provision of information relating to advertising is without prejudice to the application of the relevant provisions of Regulation (EU) 2016/679, in particular those regarding the right to object, automated individual decision-making, including profiling, and specifically the need to obtain consent of the data subject prior to the processing of personal data for targeted advertising.* |

### Article 40(3)

Article 40(3) enables the European Commission to request explanations about the design, logic, functioning, and testing of algorithmic systems like recommender systems from very large online platforms and search engines.

Initial codes proposed were:

- Dimension: Process, Business
  - Explaining the algorithmic system relates to how it functions.
  - Design and testing relates to organizational practices.
- Stakeholder: Auditor, Provider. Commission can request explanations from platforms/search engines.

Further analysis yielded:

- Dimension: Process. Focus seems to be the algorithmic system itself.
- Stakeholder: Auditor, Provider. Commission compels platforms for explanations.

After discussion, it was agreed that Business practices were also relevant given the mention of design and testing.

The final consensus was:

- Dimension: Process, Business
- Stakeholder: Auditor, Provider

Recital 141 further discusses the Commission's investigatory powers.

### Article 69(2)(d),(5)

This article empowers the European Commission to compel explanations about algorithms, data practices, and business conduct from very large online platforms during inspections.

Initial codes proposed were:

- Dimension: Data, Process, Business. Covers algorithms, data handling, and business practices.
- Stakeholder: Auditor, Provider. Commission requests explanations from platforms.

Further analysis yielded the same codes:

- Dimension: Data, Process, Business

| **Art.40(3)** | **Recital 141** |
|---|---|
| *Data access and scrutiny*<br>*1. Providers of very large online platforms or of very large online search engines shall provide the Digital Services Coordinator of establishment or the Commission, at their reasoned request and within a reasonable period specified in that request, access to data that are necessary to monitor and assess compliance with this Regulation.*<br>*3. For the purposes of paragraph 1, providers of very large online platforms or of very large online search engines shall, at the request of either the Digital Service Coordinator of establishment or of the Commission,* ***explain the design, the logic, the functioning and the testing of their algorithmic systems, including their recommender systems.*** | *The Commission should be able to request information necessary for the purpose of ensuring the effective implementation of and compliance with the obligations laid down in this Regulation, throughout the Union. In particular, the Commission should have access to any relevant documents, data and information necessary to open and conduct investigations and to monitor the compliance with the relevant obligations laid down in this Regulation, irrespective of who possesses the documents, data or information in question, and regardless of their form or format, their storage medium, or the precise place where they are stored. The Commission should be able to directly require by means of a duly substantiated request for information that the provider of the very large online platform or of the very large online search engine concerned as well as any other natural or legal persons acting for purposes related to their trade, business, craft or profession that may be reasonably aware of information relating to the suspected infringement or the infringement, as applicable, provide any relevant evidence, data and information.* ***In addition, the Commission should be able to request any relevant information from any public authority, body or agency within the Member State for the purpose of this Regulation. The Commission should be able to require access to, and explanations by means of exercise of investigatory powers, such as requests for information or interviews, relating to documents, data, information, data-bases and algorithms of relevant persons, and to interview, with their consent, any natural or legal persons who may be in possession of useful information and to record the statements made by any technical means.*** *The Commission should also be empowered to undertake such inspections as are necessary to enforce the relevant provisions of this Regulation. Those investigatory powers aim to complement the Commission's possibility to ask Digital Services Coordinators and other Member States' authorities for assistance, for instance by providing information or in the exercise of those powers.* |

- Stakeholder: Auditor, Provider

The key points of alignment were:

- The articles broadly enable transparency across data, systems, and organizational practices.
- Explanations are provided by platforms to the Commission during audits.

Data was considered a secondary explainability target since data-handling is contemplated, yet the focus seems more on auditing overall organizational algorithmic systems and procedures.

The final consensus codes were:

- Dimension: Data, Process, Business

- Stakeholder: Auditor, Provider

Recital 146 further discusses disclosing information to platforms subject to Commission decisions.

| **Art.69(2)(d),(5)** | **Recital 146** |
|---|---|
| *Power to conduct inspections 1. In order to carry out the tasks assigned to it under this Section, the Commission may conduct all necessary inspections at the premises of the provider of the very large online platform or of the very large online search engine concerned or of another person referred to in Article 67(1).*<br>*2. The officials and other accompanying persons authorised by the Commission to conduct an inspection shall be empowered to:*<br>*(d)* ***require the provider of the very large online platform or of the very large online search engine or the other person concerned to provide access to and explanations on its organisation, functioning, IT system, algorithms, data-handling and business practices and to record or document the explanations given.***<br>*5. During inspections, the officials and other accompanying persons authorised by the Commission, the auditors and experts appointed by the Commission, the Digital Services Coordinator or the other competent authorities of the Member State in the territory of which the inspection is conducted may* ***require the provider of the very large online platform or of the very large online search engine or other person concerned to provide explanations on its organisation, functioning, IT system, algorithms, data-handling and business conducts, and may address questions to its key personnel.*** | *The provider of the very large online platform or of the very large online search engine concerned and other persons subject to the exercise of the Commission's powers whose interests may be affected by a decision should be given the opportunity of submitting their observations beforehand, and the decisions taken should be widely publicised. While ensuring the rights of defence of the parties concerned, in particular, the right of access to the file, it is essential that confidential information be protected. Furthermore, while respecting the confidentiality of the information,* ***the Commission should ensure that any information relied on for the purpose of its decision is disclosed to an extent that allows the addressee of the decision to understand the facts and considerations that led up to the decision.*** |

### Article 72(1)

Article 72(1) empowers the European Commission to compel explanations and access related to algorithms and databases from very large online platforms and search engines to monitor compliance.

Initial codes proposed were:

- Dimension: Data, Process. Covers access to databases and algorithms.
- Stakeholder: Auditor, Provider. Commission can request explanations from platforms/search engines.

Further analysis yielded:

- Dimension: Data, Process, Business. Design and testing relates to organizational practices.
- Stakeholder: Auditor, Provider. Commission compels platforms for explanations.

After discussion, it was agreed that Business should be included given the mention of assessing implementation and compliance, which relates to organizational practices.

The final consensus was:

- Dimension: Data, Process, Business
- Stakeholder: Auditor, Provider

Recital 93 further discusses audits and the Commission's monitoring powers requiring transparency from platforms.

| **Art.72(1)** | **Recital 93** |
|---|---|
| *Monitoring actions*<br>*1. For the purposes of carrying out the tasks assigned to it under this Section, the Commission may take the necessary actions to monitor the effective implementation and compliance with this Regulation by providers of the very large online platform and of the very large online search engines. The Commission may order them to provide* ***access to, and explanations relating to, its databases and algorithms.*** *Such actions may include, imposing an obligation on the provider of the very large online platform or of the very large online search engine to retain all documents deemed to be necessary to assess the implementation of and compliance with the obligations under this Regulation.* | ***The audit report should be substantiated, in order to give a meaningful account of the activities undertaken and the conclusions reached. It should help inform, and where appropriate suggest improvements to the measures taken by the providers of the very large online platform and of the very large online search engine to comply with their obligations under this Regulation.*** *The audit report should be transmitted to the Digital Services Coordinator of establishment, the Commission and the Board following the receipt of the audit report. Providers should also transmit upon completion without undue delay each of the reports on the risk assessment and the mitigation measures, as well as the audit implementation report of the provider of the very large online platform or of the very large online search engine showing how they have addressed the audit's recommendations. The audit report should include an audit opinion based on the conclusions drawn from the audit evidence obtained.* [...] ***Where the audit opinion could not reach a conclusion for specific elements that fall within the scope of the audit, an explanation of reasons for the failure to reach such a conclusion should be included in the audit opinion. Where applicable, the report should include a description of specific elements that could not be audited, and an explanation of why these could not be audited.*** |

## A.2.9 Digital Markets Act – 32022R1925 - EN - EUR-Lex

### Article 21(1),(2)

Initial codes proposed for Article 21(1),(2) were:

- Dimension: Data, Process. The Article enables access to algorithms, data, testing information.
- Stakeholder: Auditor, Provider. Commission requests information, gatekeepers provide it.

Further analysis yielded:

- Dimension: Data, Process, Business. Agreement on data and algorithms. Testing info could relate to business practices.
- Stakeholder: Auditor, Provider, User. Gatekeepers ultimately provide transparency to users.

After discussion, it was agreed that testing processes involve organizational practices, so Business is relevant but secondary, since the most relevant dimensions appear to be data, algorithms, and testing procedures.

Final consensus:

- Dimension: Data, Process, Business
- Stakeholder: Auditor, Provider

Recital 81 further emphasizes Commission access to any necessary information related to the AI systems.

| **Art.21(1),(2)** | **Recital 81** |
|---|---|
| *Requests for information*<br>*1. In order to carry out its duties under this Regulation, the Commission may, by simple request or by decision, require from undertakings and associations of undertakings to provide all necessary information. The Commission may also, by simple request or by decision,* ***require access to any data and algorithms of undertakings and information about testing, as well as requesting explanations of them.***<br>*2. When sending a simple request for information to an undertaking or association of undertakings, the Commission shall state the legal basis and purpose of the request, specify what information is required and fix the time limit withiN which the information is to be provided, as well as the fines provided for in Article 30 applicable for supplying incomplete, incorrect or misleading information or explanations.* | *The Commission should be empowered to request information necessary for the purpose of this Regulation. In particular, the Commission should have access to any relevant documents, data, database, algorithm and information necessary to open and conduct investigations and to monitor the compliance with the obligations laid down in this Regulation, irrespective of who possesses such information, and regardless of their form or format, their storage medium, or the place where they are stored.* |

### Article 23

Article 23 empowers the European Commission to compel explanations about algorithms, data practices, and business conduct from undertakings during inspections.

Initial codes proposed were:

- Dimension: Data, Process, Business. Covers data handling, IT systems, algorithms and business practices.
- Stakeholder: Auditor, Provider. Commission can require explanations from undertakings.

Further analysis yielded the same codes:

- Dimension: Data, Process, Business
- Stakeholder: Auditor, Provider

The key points of alignment were:

- The article enables transparency across data, systems, and organizational practices.
- Explanations are provided by undertakings to the Commission during audits.

Since the codes were independently aligned, no discussion was needed.
The final consensus codes were:

- Dimension: Business, Data, Process
- Stakeholder: Auditor, Provider

| **Art.23** | **Recital 83** |
|---|---|
| *Powers to conduct inspections*<br>1. *In order to carry out its duties under this Regulation, the Commission may conduct all necessary inspections of an undertaking or association of undertakings.*<br>2. *The officials and other accompanying persons authorised by the Commission to conduct an inspection are empowered to:*<br>1. *enter any premises, land and means of transport of undertakings and associations of undertakings;*<br>2. *examine the books and other records related to the business, irrespective of the medium on which they are stored;*<br>3. *take or obtain in any form copies of or extracts from such books or records;*<br>4. ***require the undertaking or association of undertakings to provide access to and explanations on its organisation, functioning, IT system, algorithms, data-handling and business practices and to record or document the explanations given by any technical means;***<br>5. *seal any business premises and books or records for the duration of, and to the extent necessary for, the inspection*;<br>6. *ask any representative or member of staff of the undertaking or association of undertakings for explanations of facts or documents relating to the subject-matter and purpose of the inspection, and to record the answers by any technical means.*<br>4. *During inspections the Commission, auditors or experts appointed by it and the national competent authority of the Member State, enforcing the rules referred to in Article 1(6) in whose territory the inspection is to be conducted* ***may require the undertaking or association of undertakings to provide access to and explanations on its organisation, functioning, IT system, algorithms, data-handling and business conducts****. The Commission and auditors or experts appointed by it and the national competent authority of the Member State, enforcing the rules referred to in Article 1(6) in whose territory the inspection is to be conducted may address questions to any representative or member of staff.* | *The Commission should also be empowered to conduct inspections of any undertaking or association of undertakings and to interview any persons who could be in possession of useful information and to record the statements made.* |

# APPENDIX B

# XAI ETHICS CLASSIFICATIONS IN CHAPTER 5

## B.1 XAI Ethics Classification

Recognizing the diverse ways in which ethical considerations can be integrated into XAI research, we have developed a systematic classification scheme to assess the depth and quality of ethical engagement across the analyzed literature. This classification protocol serves as a guide for researchers to examine the content and focus of each paper, and subsequently assign it to one of five categories (A-E). The categories are differentiated based on three key dimensions: (i) the depth of ethical discussion, (ii) the application of specific ethical theories or frameworks, and (iii) the overall emphasis on ethical issues in relation to XAI. The resulting classifications aims to reveal the prevalence of ethical discussions, alongside the extent to which these discussions are substantive, grounded in normative theories, and explicitly linked to the design and development of XAI tools and techniques.

To facilitate the assignment of each classification category (A-E), a rating system based on quantitative thresholds is established. These thresholds are based on key criteria and provide more precise and objective classification. In this structured classification scheme, the depth of ethical discussion is evaluated through a Likert scale (Step 2), while the presence and application of ethical theories or frameworks are assessed separately (Step 3). The overall focus of the paper on ethical issues in XAI is also rated on a Likert scale (Step 4). These ratings are then summed to determine the quantitative thresholds for each category, as defined in Table B.2. By combining a rigorous protocol with quantitative thresholds and illustrative examples reported

| Step | Action | Response | Guidance |
|---|---|---|---|
| **1** | Identify explicit discussions of ethics in the context of XAI in the paper. | Yes/No | If "Yes," the paper can be included in the classification. |
| **2** | Evaluate the depth of ethical discussions. This should take into account the complexity, thoroughness, and sophistication of the ethical argumentation. | Likert scale (1-5) | A rating of 1 indicates ethical considerations are only briefly mentioned, while a rating of 5 signifies an extensive ethical discussion with deep analysis of ethical principles in relation to XAI. |
| **3** | Assess if the paper refers to any specific ethical theories or frameworks, and how they are applied to XAI. | Yes/No + Description | If "Yes," specify the theory or framework and its application. |
| **4** | Determine the main focus of the paper. This can be identified by understanding the research question, objectives, and the contribution the paper is making to the field of XAI. | Likert scale (1-5) | A rating of 1 indicates the paper is not focused on ethics, while a rating of 5 signifies an extreme focus on ethics in relation to XAI. |
| **5** | Assign the most appropriate category (A-E) to the paper based on steps 1-4 (Refer to Table 2a below) | A, B, C, D, E | The assigned category should reflect the overall depth in ethical discussion, the application of ethical theories/frameworks, and the focus on ethical issues in XAI. |

**Table B.1:** Classification Protocol

in B.1.1, this classification approach aims to promote consistency, objectivity, and reproducibility in assessing the ethical dimensions of XAI research.

### B.1.1 Justification and Structure of the Classification Scheme (A-E)

To further clarify the distinctions between categories, we provide illustrative examples from the analyzed literature:

- **Category A:** A paper that merely states e.g., “Ethical issues are important in XAI development" without any further analysis would fall into this category.
- **Category B:** A paper discussing the need for transparency and fairness in XAI systems, but not delving into a deeper examination of these ethical principles, would be classified as Category B.

| Cat. | Definition | Quantitative Threshold | Guidance |
|---|---|---|---|
| **A** | Papers that merely mention ethics or ethical values but do not engage in any ethical analysis. | Ethical Discussion Score: 1-2 | The paper makes only passing reference to ethics or ethical values, with no meaningful analysis or discussion. |
| **B** | Papers that discuss ethical values or principles in the context of XAI without providing a thorough ethical analysis. | Ethical Discussion Score: 3-4 | The paper includes a discussion of ethics, but the analysis is largely surface level, lacking in depth and sophistication. |
| **C** | Papers that present a systematic ethical analysis, but the ethical considerations are not explicitly linked to the design or development of XAI tools. | Ethical Discussion Score: 4-5<br>Ethical Theory Mentioned: Yes<br>Primary Focus Score: 1-3 | The paper engages in a systematic ethical analysis, but does not link these ethical considerations to XAI design or development. |
| **D** | Papers that propose XAI tools or techniques that are informed by ethical considerations, but the connection between the ethical principles and the XAI solutions is not thoroughly substantiated. | Ethical Discussion Score: 4-5<br>Ethical Theory Mentioned: Yes<br>Primary Focus Score: 3-4 | The paper proposes XAI tools or techniques that are informed by ethical considerations, but the connection between the ethical principles and the XAI solutions is not thoroughly substantiated. |
| **E** | Papers that explicitly integrate ethical considerations into the design and development of XAI tools, and provide a comprehensive and rigorous ethical analysis of the proposed solutions. | Ethical Discussion Score: 5<br>Ethical Theory Mentioned: Yes<br>Primary Focus Score: 5 | The paper explicitly integrates ethical considerations into the design and development of XAI tools, and provides a comprehensive and rigorous ethical analysis of the proposed solutions. |

**Table B.2:** Quantitative Thresholds for Classification Categories

- **Category C:** A paper that systematically analyzes the application of deontological ethics (e.g., Kantian ethics) to XAI, but does not explicitly link this analysis to the design or development of XAI tools, would be considered Category C.

- **Category D:** A paper proposing an XAI technique for enhancing fairness, citing ethical principles of non-discrimination, but without thoroughly substantiating the connection between the proposed technique and the ethical principles, would fall under Category D.

- **Category E:** A paper that explicitly grounds the development of an XAI tool in the ethical framework of care ethics, providing a rigorous analysis of how

the tool's design and implementation uphold the principles of attentiveness, responsibility, competence, and responsiveness, would be classified as Category E.

While the classification scheme aims to capture distinct levels of ethical integration, it is important to acknowledge its inherent limitations and potential biases. The assessment of the depth of ethical discussion and the determination of a paper's primary focus inevitably involve some degree of subjectivity, despite the efforts to establish clear criteria and quantitative thresholds. Additionally, annotator biases may persist despite the training and conflict resolution measures employed.

### B.1.2 Research Queries on Scopus

The research queries employed in this study were carefully crafted to encompass the diverse ethical considerations relevant to the XAI field. The selection of search terms was grounded in the key ethical theories, principles, and debates identified as pertinent to the design, development, and deployment of XAI systems. Our primary research question aimed to assess the extent and depth of ethical discussions within XAI research and the application of ethical theories or frameworks in this domain. To address this question comprehensively, we adopted a two-pronged approach in constructing our search queries:

- **Foundational Ethical Theories:** We incorporated terms related to the major normative ethical theories, such as deontology, consequentialism, virtue ethics, and care ethics. These theories provide the philosophical underpinnings for many of the ethical principles and frameworks discussed in the context of AI and XAI.
- **Applied Ethics in XAI:** We included terms specific to ethical principles and concepts relevant to XAI, such as transparency, accountability, fairness, and responsible AI design. These principles capture the unique ethical challenges and considerations that arise in the development and deployment of explainable AI systems.

The following Scopus search queries were used, reflecting this comprehensive approach:

```
TITLE-ABS-KEY ( "explainable AI" OR "interpretable AI" AND "
    ↪ ethical theories" AND "application" ) = 0
TITLE-ABS-KEY("Explainable AI" OR "XAI" OR "interpretable
    ↪ machine learning" OR "interpretability" OR "AI
    ↪ explainability") AND TITLE-ABS-KEY("ethics" OR "ethical"
    ↪  OR "moral" OR "morality" ) = 409
TITLE-ABS-KEY("Explainable AI" OR "XAI" OR "interpretable
    ↪ machine learning" OR "interpretability" OR "AI
    ↪ explainability") AND TITLE-ABS-KEY("deontology" OR "
    ↪ consequentialism" OR "virtue ethics" OR "care ethics" OR
    ↪  "ethics of care" OR "utilitarianism" OR "rights-based
    ↪ ethics" OR "contractualism" OR "social contract theory"
    ↪ OR "relational ethics" OR "distributive justice") = 4
TITLE-ABS-KEY("Explainable AI" OR "XAI" OR "interpretable
    ↪ machine learning" OR "interpretability" OR "AI
    ↪ explainability") AND TITLE-ABS-KEY("ethics" OR "ethical"
    ↪  OR "moral" OR "morality" AND "responsible AI" OR "
    ↪ ethical design" OR "ethical impact assessment") = 25
```

### B.1.3 Results

In the primary phase of our bibliometric study, an initial pool of 410 research papers was established. Following the application of our predefined inclusion criteria, we subsequently eliminated 173 of these articles, leaving a sample of 237 papers for further review. Within this remaining pool, each paper was thoroughly examined, with both abstract and body text read and analyzed. Prior to the final classification process, an additional elimination of papers deemed as not relevant was undertaken. These were primarily research articles that emerged as false positives in our methodology – papers not directly applicable to our study focus. These included a total of 143 papers that treated the subjects of XAI or ethical considerations independently, without a focus on their intersection. This category also encompassed 17 survey articles that were identified within our pool. The entire process of filtering and categorization is visually depicted in Figure B.1.

#### Overview of Paper Distribution

Our multi-stage filtering process resulted in a final pool of 77 research papers for in-depth analysis. These papers were classified according to our pre-established

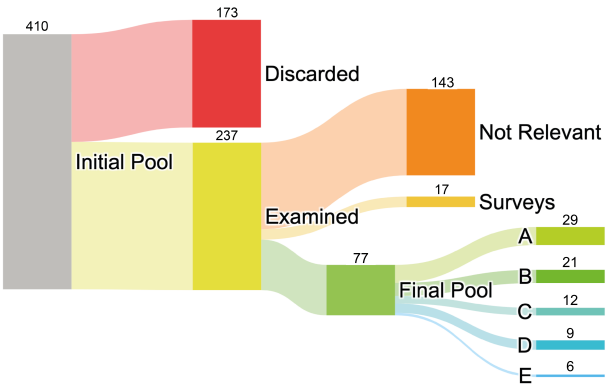


**Figure B.1:** Decision tree illustrating the distribution of papers at distinct stages of the process.

five-tiered ranking system (A-E), which assessed the relevance and depth of ethical engagement in the context of XAI research.

***Distribution across Categories*** – The distribution of papers across the five categories comprised 29 papers (37.66% of the pool) occurrences in Category **A**; 21 – 27.27% in Category **B**; 12 – 15.58% in Category **C**; 9 – 11.69% in Category **D**; 6 – 7.79% in Category **E**. Notably, over 60% of the papers fell into categories A and B, indicating a relatively superficial engagement with ethical considerations in a significant portion of XAI research. In contrast, only about 20% of the papers (categories D and E) demonstrated a deeper integration of ethical analysis into the design and development of XAI systems. Out of the 77 papers in the list, 39 (50.6%) were published in conference proceedings, 34 (44.2%) in journals, and 4 (5.2%) in workshops or other publication types. This distribution highlights the importance of both conferences and journals in advancing research on ethics in XAI.

***Key Publication Venues*** – Several conferences and journals have emerged as key outlets for research on ethics in XAI, as reported in Table B.3. The most prominent venue in the list is *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), with 9 papers. This is followed by the conference *AIES* (AAAI/ACM Conference on AI, Ethics, and Society) with 4 papers; *Philosophy and Technology* with 2 papers and *Ethics and Information Technology* with 3 papers. Other notable venues include *Minds and Machines* with 2 papers, *IEEE International Conference on Fuzzy Systems* with 2 papers, and *Advances in Intelligent Systems and Computing* with 2 papers. These venues highlights the multidisciplinary nature of research on ethics in XAI.

| Key Publication Venues | References |
|---|---|
| Lecture Notes in Computer Science (including subseries Lecture Notes in AI and Lecture Notes in Bioinformatics) | [50, 63, 105, 179, 283, 310, 319, 476, 511] |
| AIES (AAAI/ACM Conference on AI, Ethics, and Society) | [266,436,452,513] |
| Ethics and Information Technology | [225, 239, 458] |
| FAccT (ex-FAT) | [188, 232] |
| Philosophy and Technology | [47, 199] |
| Minds and Machines | [344, 405] |
| IEEE International Conference on Fuzzy Systems | [22, 196] |
| Advances in Intelligent Systems and Computing | [23, 168] |
| Frontiers in Artificial Intelligence and Applications | [139] |

**Table B.3:** Key Publication Venues and References

In terms of disciplinary focus of publication venues, we report an excerpt of distribution across the categories that most consistenly engaged with ethical consideration in Table B.4.

The full complete list seen a majority of papers (36%) coming from Computer Science outlets; Medicine/Healthcare venues account for the 32%; Ethics & Society outlets account for the 16%; Law venues are represented as the 12%; and Business/-Management as the 4%. This distribution showcases the multidisciplinary nature of research on ethics in XAI, with significant contributions from computer science, medicine/healthcare, ethics & society, law, and business/management outlets. The strong representation of computer science and medicine/healthcare venues highlights technical and domain-specific considerations in the development and application of ethical XAI systems.

| Domain | Papers |
|---|---|
| Computer Science | [202], [47], [121], [347], [168], [105], [283], [139], [476] |
| Ethics & Society | [147], [270], [344], [232] |
| Medicine/Healthcare | [309], [333], [225], [239], [197], [199], [472], [25] |
| Law | [175], [429], [223] |
| Business/Management | [452] |

**Table B.4:** Disciplinary Domains and References among **C**, **D**, and **E** categories.

### Depth and Extent of Ethical Discussions

The majority of XAI articles in categories A-B make reference to ethical considerations regarding AI exclusively in the abstract or in the introduction section, without further

developing the discussion. Ethics is often presented as a motivation for the work or used to contextualize the proposed XAI methods within the landscape of real-world applications. In practice, applications and ethical implications are almost always mentioned together, alongside legal issues. Furthermore, these considerations are typically used to introduce the term XAI in general as an AI ethics principle, rather than being concretely connected to the specific proposed method. Papers classified under categories C, D, and E demonstrated varying, yet more substantiated, levels of engagement with ethical theories and frameworks. A closer examination of the papers in each category reveals distinct patterns in the depth and quality of ethical engagement:

- **Category A** papers, which constituted the largest group (37.66%), typically mentioned ethics or ethical values in passing without engaging in any substantive ethical analysis. Many of these papers referred to ethics in the abstract or introduction as a general motivation for the work, but failed to connect these considerations to the specific XAI methods or applications being proposed.

- **Category B** papers (27.27%) went a step further by discussing ethical principles or values in the context of XAI, but still lacked a thorough or systematic ethical analysis. These papers often highlighted the importance of ethical considerations such as transparency, accountability, or fairness, but did not delve into the nuances of how these principles might be operationalized or navigated in practice.

- **Category C** papers (15.58%), such as [121,147,175,197,225,270,290,309,333,344,347, 487], present ethical analyses but do not explicitly link the ethical considerations to the design or development of specific XAI tools. Instead, they focus on critiquing existing approaches, highlighting ethical challenges, or proposing conceptual frameworks and guidelines for addressing ethical issues in XAI.

- **Category D** papers (11.69%) began to bridge this gap by proposing XAI tools or techniques that were informed by ethical considerations. Examples such [47, 70, 105, 168, 202, 232, 239, 283, 429, 476], propose XAI tools or techniques informed by ethical considerations but do not thoroughly substantiate the connection between the ethical principles and the proposed solutions. These papers often focus on specific aspects of explainable AI, such as generating explanations for moral judgments [283], classifying AI crimes [429], or designing human-agent collaboration protocols [472]. However, the connection between the ethical principles invoked and the specific XAI solutions proposed was not always thoroughly substantiated or explored in depth.

- **Category E** papers, while representing the smallest proportion (7.79%), offered the most comprehensive and rigorous integration of ethical considerations into the design

and development of XAI systems. Papers such as [25, 139, 199, 223, 452, 472], explicitly integrate ethical considerations into the design and development of XAI tools and provide comprehensive ethical analyses of the proposed solutions. These papers engage more deeply with ethical theories and frameworks, using them to guide the design and evaluation of explainable AI systems. For example, [25] conducts an ethical assessment of explainability in AI-based clinical decision support using the "Principles of Biomedical Ethics," while [452] proposes using the capability approach to provide ethical standards for algorithmic recourse. Notably, papers from philosophical, social science, and interdisciplinary backgrounds (e.g., [25]) often provide more extensive engagement with ethical theories and frameworks compared to papers from purely technical domains.

Across the papers reviewed, a diverse range of ethical theories and frameworks are applied to analyze the role of explainability in AI systems. Certain ethical theories and principles emerged as more prominent than others. *Consequentialism, Deontological Ethics, Virtue Ethics*) are relatively present. In Table B.5, we report a list of works that refer to them more or less explicitly. Explicit mentions are also to be found to the "Principles of Biomedical Ethics" by Beauchamp & Childress (autonomy, beneficence, nonmaleficence, and justice) e.g., used as an analytical framework in [25] to assess the ethical implications of explainability in AI-based clinical decision support systems. Similarly, [199] builds on the notion of "explicability" proposed by Floridi et al., which combines the demands for intelligibility and accountability of AI systems, to argue for the ethical and epistemological utility of explainable AI in medicine.

| **Ethical Theories** | **#** | **Refs.** |
|---|---|---|
| *Consequentialism* | 13 | [46, 50, 175, 196, 199, 225, 239, 245, 319, 333, 344, 405, 429] |
| *Deontological ethics* | 6 | [42, 46, 50, 283, 344, 429] |
| *Virtue ethics* | 10 | [46, 50, 105, 121, 282, 283, 319, 333, 429, 511] |

**Table B.5:** References of papers that engage in a discourse regarding major ethical theories presented (not necessarily just one).

Several papers draw upon philosophical concepts and frameworks to examine the ethical dimensions of explainable AI. [344] explores the attitudinal tensions between explainability and trust in AI decision support tools, discussing the incompatible deliberative and unquestioning attitudes required for each. [232] critiques the use of counterfactuals in algorithmic fairness and explainability, arguing that social categories may not admit counterfactual manipulation and proposing tenets for using counterfactuals in machine learning. [223] introduces the concept of "denunciatory power" as an ethical desideratum for AI explanations, measuring their ability to reveal unethical decisions or behavior. [105] employs the Value Sensitive Design (VSD)

method to facilitate transparency and the realization of ethical principles in AI and digital systems design. [472] proposes three team design patterns with varying levels of agent autonomy and human involvement to enable moral decision-making in human-agent teams. Other papers engage with various ethical principles and concepts, such as informed consent, shared decision-making, accountability, fairness, and transparency [225, 239, 283, 452].

## B.2 Taxonomy of Technical and Sociotechnical Risks in XAI

We performed a research literature retrieval grounded on concerns and vulnerabilities of XAI, from where we identified key technical risks. This preliminary analysis constituted the bedrock from which we departed our thematic analysis. As a second step, our search strategy through citation chaining and snowballing incorporated diverse disciplinary perspectives, including computer science, cognitive science, psychology, law, ethics, sociology, and others, ensuring a comprehensive view of the contextual risks associated with explanations in AI. This approach was inspired by social sciences studies informing the field of XAI [284, 285, 325, 501]. This allowed us to garner a deeper understanding of how explanations function in non-AI contexts, enriching our understanding of potential risks when these concepts are transposed into the XAI domain.

**Research Retrieval & Filtering** – We began targeting various the Scopus academic database and then expanding to other peer-reviewed sources such as ACM Digital Library and IEEE Explore. For search strings, keywords or concepts such as *explainable*, *XAI*, *interpretable ML* were incorporated with terms as *vulnerabilities*, *adversarial attacks*, *robustness*, *data poisoning* and others. Terms were chosen based on our prior knowledge of common challenges and threats faced by AI systems in general and XAI systems in particular. The departing Scopus queries were:

1. Query (1) targeted technical risks related to the robustness of XAI methods, including their vulnerability to adversarial attacks, model manipulation, and input perturbations[1].

2. Query (2) focused on fairness risks in XAI, covering topics such as algorithmic bias, discrimination, disparate impact, and various fairness metrics and

[1] (1) TITLE-ABS-KEY((“explainable AI” OR “XAI” OR “interpretable machine learning”) AND (“robustness” OR “adversarial attacks” OR “adversarial examples” OR “adversarial perturbations” OR “model manipulation” OR “saliency maps” OR “counterfactual explanations” OR “concept activation vectors” OR “input perturbations”) AND (“risks” OR “vulnerabilities” OR “challenges” OR “issues”))

constraints[2].

3. Query (3) addressed privacy and security risks associated with XAI, including information leakage, model inversion attacks, membership inference attacks, model extraction, and risks to intellectual property[3].

To ensure a comprehensive search, we also included synonyms and related terms for each keyword. For example, when searching for *adversarial attacks*, we also used terms like *adversarial examples*, *adversarial perturbations*, and *adversarial manipulations*. This approach helped capture a wider range of relevant literature that may use slightly different terminology to describe similar concepts.

**Selection Criteria & Analysis** – To ensure the relevance and quality of the articles included in our analysis, we included papers: (I°.) Published in a peer-reviewed journal, conference proceedings, or book chapters; (II°.) Focused on explainable AI from a perspective informed by risk assessment, associated vulnerabilities, or AI ethics frameworks; (III°.) Presented a theoretical or empirical analysis of risks related to XAI explanations, system architectures, or data; (IV°.) Written in English.

In addition to the structured search of XAI-specific literature, from our paper pool we expanded to similar works through citation chaining and snowballing incorporated diverse disciplinary perspectives, including computer science, cognitive science, psychology, law, ethics, sociology, and others. We deliberately included papers from non-XAI/AI contexts, particularly from the period before the establishment of the XAI program by DARPA in 2016 [180, 366]. This decision was motivated by the recognition that the study of explanations has a long and rich history in fields such as psychology, cognitive science, philosophy, and human-computer interaction – e.g., [82, 192, 198, 238, 292, 415, 416, 461, 502]. By drawing from this diverse body of knowledge, we aimed to gain a more comprehensive understanding of the potential risks and challenges associated with explanations in human communication, and to identify foundational concepts and theories that have shaped the current understanding of explainability in AI [91].

[2](2) TITLE-ABS-KEY((“explainable AI” OR “XAI” OR “interpretable machine learning”) AND (“fairness” OR “bias” OR “discrimination” OR “disparate impact” OR “demographic parity” OR “equal opportunity” OR “algorithmic fairness” OR “fairness metrics” OR “fairness constraints” OR “fairness-aware learning”) AND (“risks” OR “vulnerabilities” OR “challenges”))

[3](3) TITLE-ABS-KEY((“explainable AI” OR “XAI” OR “interpretable machine learning”) AND (“privacy” OR “security” OR “information leakage” OR “model inversion” OR “membership inference” OR “model extraction” OR “gradient leakage” OR “intellectual property” OR “trade secrets” OR “privacy-preserving” OR “secure multiparty computation”) AND (“risks” OR “vulnerabilities” OR “challenges”))"

**Data extraction and analysis** – In analyzing this collection of papers, we adopted an iterative and reflexive process. We derived key themes directly from the literature and honed through continuous comparison with our expanding dataset[4]. In particular, the thematic analysis was conducted in six phases following the guidelines proposed by [58]:

1. *Familiarization with the data*: The researchers read the selected papers to gain an understanding of the content.

2. *Generating initial codes*: Each researcher independently coded a subset of the papers, identifying initial themes and patterns related to XAI risks.

3. *Searching for themes*: Through an iterative process of discussion and refinement, the researchers developed a preliminary set of themes and subthemes that captured the key risks associated with XAI systems.

4. *Reviewing themes*: The researchers independently reviewed the preliminary themes and subthemes, checking their coherence and consistency against the coded data and the original papers. The researchers then met to discuss their findings and refine the themes and subthemes accordingly.

5. *Defining and naming themes*: The researchers collaboratively defined and named the final set of themes and subthemes, ensuring that each theme captured a distinct and meaningful aspect of XAI risks.

We clarify that this partitioning into themes and subthemes is inherently interpretive and adaptive. We acknowledge that due to the complexity of the field and the variable lexicon used across the literature, certain papers may resonate with multiple subthemes or themes.

## B.3 Categorization of Risks in XAI Systems

We developed a taxonomy categorizing the identified risks into two primary domains: *Technical Risks*, related to the data and models of XAI systems, and *Contextual Risks*, associated with the informativeness and reception of XAI explanations. The interested

[4]The thematic analysis was conducted by a team of three researchers with diverse expertise in XAI, AI ethics, and qualitative research methods. This interdisciplinary team composition ensured a comprehensive and rigorous analysis of the data. Researcher 1 (R1) has a background in computer science and XAI, with extensive experience in developing and evaluating XAI methods. Researcher 2 (R2) specializes in AI ethics and has published on the social and ethical implications of AI systems. Researcher 3 (R3) is an expert in qualitative research methods.

reader can visualize the taxonomy in table B.3 in Table B.3 below. Risks reported are to be considered as not mutually excluding[5].

| Category | Subcategory | References |
|---|---|---|
| ***Technical*** | | |
| **Robustness** | Attacks on saliency-based explanation methods | [10, 24, 296, 402, 436, 456, 503, 509, 512] |
| | Manipulation of counterfactual explanations | [235, 236, 240, 261, 327, 435, 447, 484, 486] |
| | Attacks on concept-based explanation methods | [61, 170, 243, 434] |
| | Adversarial perturbations affecting explanations | [43, 314, 341, 440, 456, 510] |
| | Explanation-aware backdoors | [349] |
| | Debugging challenges | [11, 12, 95] |
| | Transferability of adversarial attacks | [265, 433] |
| **Fairness** | Fairwashing | [19, 20] |
| | Biased sampling | [161, 264] |
| | Adversarial poisoning | [314, 440] |
| | Manipulation of post-hoc explanations | [108, 264, 320] |
| | Explanation disparity risks | [41, 95] |
| **Evaluation** | Dependence on model assumptions | [34, 348] |
| | Evaluation manipulation and deception | [11, 492] |
| | Robustness-explainability trade-off | [14, 348, 413] |
| | Reliability of interpretation methods | [203, 209, 244, 459] |
| ***Contextual*** | | |
| **Security** | Privacy vulnerabilities | [110, 286, 391, 427] |
| | Instrumentalization | [16, 113, 263, 321, 368, 384, 407, 497] |
| **Accountability** | Traceability of explanation design | [83, 495] |
| | Appraising explainers | [260, 507] |
| | Explainer's overconfidence | [148, 149, 258, 259, 505] |

[5] We decided to arbitrarily adopt a categorization that reflects both the themes of literature retrieval and filtering exposed before, as well as citation chaining. We consider thus some of these risks mutual e.g., adversarial attacks can be used to manipulate the input data of the underlying AI system, which in turn can affect the fairness of the explanations generated by the XAI system; biased sociotechnical explanations (e.g., essentialism) might be used to justify unfair data distributions; technical privacy risks easily overlap with gaming opportunities, etc.

| | | |
|---|---|---|
| **Heuristics & Reception** | Cognitive heuristics | [226, 464] |
| | Implications of language and semantic framing | [227, 279] |
| | Cognitive biases | [346, 411, 462, 464] |
| **Argumentative & Logical** | Circular reasoning | [185, 490] |
| | Tautology | [315, 389, 446] |
| **Under & Over Determination** | Underdetermination | [103, 262, 277, 446] |
| | Overdetermination | [291, 488] |
| **Reification & Essentialism** | Reification | [195, 211, 267, 268, 420, 423, 478, 493] |
| | Essentialism | [104, 214, 311, 401, 409] |

**Table B.6:** Categorization of Risks

# Appendix C

# Self-Assessment Tool for XAI in Process Mining

This self-assessment tool is designed to help PM practitioners evaluate their readiness for implementing XAI techniques in their projects. By answering a series of questions related to project details, organizational context, explainability requirements, and constraints, users can generate personalized recommendations and action items to guide their XAI adoption journey. This tool is grounded in the multi-step framework for integrating XAI into PM practices, as detailed in Chapter 6.3. The framework provides a structured approach to understanding the domain, enhancing PM practices, addressing data and system integration barriers, making PM insights more accessible, and implementing and validating XAI techniques.

The self-assessment tool's structure and content are directly informed by the framework's emphasis on understanding the project context (Step 1), mapping stakeholder needs and constraints (Steps 0-2), and selecting appropriate XAI techniques and delivery strategies (Steps 3-4). The questions in the "Project Details" and "Organizational Context" sections of the self-assessment tool correspond to the domain understanding and context mapping steps of the framework, helping users identify the specific characteristics and requirements of their PM projects and organizational environments. The "Explainability Requirements" section of the self-assessment tool aligns with the framework's guidance on deconstructing stakeholder needs (Step 2) and associating appropriate explanation types and targets (Step 3). The questions in this section prompt users to consider the specific explainability needs of different stakeholders, the objectives of providing explanations, and the aspects of the PM results that require explainability. The decision logic associated with these questions

draws upon the framework's recommendations for selecting suitable XAI techniques based on the identified requirements. Finally, the "Constraints and Priorities" section of the self-assessment tool addresses the framework's emphasis on considering technical and organizational constraints, as well as balancing competing priorities, when implementing XAI solutions (Steps 4-5). The questions in this section encourage users to assess the limitations and trade-offs they may face in their specific contexts, while the decision logic provides guidance on how to navigate these challenges based on the insights from the framework.

**Project Details**

1. What are the primary objectives of your process mining project? (Select all that apply)
   - □ Process discovery
   - □ Conformance checking
   - □ Performance analysis
   - □ Predictive monitoring
   - □ Root cause analysis
   - □ Other (please specify)
2. What is the main process or domain area you are focusing on?
   - □ Healthcare
   - □ Manufacturing
   - □ Finance
   - □ Retail
   - □ Telecommunications
   - □ Other (please specify)
3. How would you describe the quality and format of your event log data?
   - □ Structured and high-quality
   - □ Structured with some quality issues
   - □ Unstructured or complex format
   - □ Incomplete or noisy data
   - □ Other (please specify)
4. What is the approximate size of your event log?
   - □ Small (less than 1,000 cases and 10,000 events)
   - □ Medium (1,000 to 10,000 cases and 10,000 to 100,000 events)
   - □ Large (more than 10,000 cases and 100,000 events)
5. Are there any specific data attributes or perspectives that are crucial for your analysis? (Select all that apply)
   - □ Timestamp

- ◻ Resource
- ◻ Cost
- ◻ Quality
- ◻ Other (please specify)

**Organizational Context**

1. What is your organization's primary industry?
   - ◻ Healthcare
   - ◻ Manufacturing
   - ◻ Finance
   - ◻ Retail
   - ◻ Telecommunications
   - ◻ Other (please specify)
2. How would you describe your organization's level of process mining maturity?
   - ◻ Beginner (limited experience, no dedicated tools or infrastructure)
   - ◻ Intermediate (some experience, basic tools, and ad-hoc projects)
   - ◻ Advanced (extensive experience, robust tools, and enterprise-wide deployment)
3. What are the key stakeholder roles involved in your process mining project? (Select all that apply)
   - ◻ Process owners
   - ◻ Business analysts
   - ◻ Data scientists
   - ◻ IT professionals
   - ◻ Domain experts
   - ◻ Other (please specify)
4. How would you characterize your organization's data governance and security policies?
   - ◻ Strict (rigorous data access controls, regular audits, and compliance checks)
   - ◻ Moderate (defined policies but inconsistent enforcement)
   - ◻ Lenient (limited policies and oversight)
5. What is the level of executive support and sponsorship for your process mining initiatives?
   - ◻ High (dedicated budget, resources, and strategic alignment)
   - ◻ Moderate (cautious support, limited resources)
   - ◻ Low (skepticism, competing priorities)

**Explainability Requirements**

1. Who are the primary stakeholders requiring explainability? (Select all that apply)
   - ◻ Process owners
   - ◻ Business analysts

- ☐ Data scientists
- ☐ End-users
- ☐ Regulators or auditors
- ☐ Other (please specify)

2. What are the main objectives of providing explanations to these stakeholders? (Select all that apply)
   - ☐ Improving trust and acceptance of process mining insights
   - ☐ Facilitating data-driven decision-making and process optimization
   - ☐ Ensuring compliance with regulatory requirements
   - ☐ Enhancing collaboration and knowledge sharing among stakeholders
   - ☐ Other (please specify)
3. What types of explanations are most important for your stakeholders? (Select all that apply)
   - ☐ Descriptive
   - ☐ Diagnostic
   - ☐ Predictive
   - ☐ Prescriptive
   - ☐ Counterfactual
4. What specific aspects of the process mining results require explainability? (Select all that apply)
   - ☐ Model predictions
   - ☐ Process deviations or anomalies
   - ☐ Performance metrics
   - ☐ Decision rules
   - ☐ Other (please specify)
5. Are there any domain-specific or organizational constraints that impact the explainability requirements? (Select all that apply)
   - ☐ Regulatory compliance
   - ☐ Contractual obligations
   - ☐ Ethical considerations
   - ☐ Other (please specify)

**Constraints and Priorities**

1. What are the main technical constraints for your PM-XAI implementation? (Select all that apply)
   - ☐ Limited data quality or availability
   - ☐ Incompatible systems or tools
   - ☐ Lack of technical expertise
   - ☐ Scalability and performance issues

- ☐ Other (please specify)

2. What are the key organizational constraints or considerations? (Select all that apply)
   - ☐ Limited resources (budget, time, personnel)
   - ☐ Resistance to change
   - ☐ Complex stakeholder relationships
   - ☐ Lack of governance or standardization
   - ☐ Other (please specify)
3. What are your top priorities for the PM-XAI project? (Select your top 3)
   - ☐ Accuracy and reliability of explanations
   - ☐ Interpretability and ease of understanding
   - ☐ Compliance with regulations and standards
   - ☐ Speed and efficiency of implementation
   - ☐ Scalability and future-proofing
   - ☐ Other (please specify)
4. How would you rank the following explainability trade-offs for your project?
   1. Transparency vs. Confidentiality
      - ☐ High transparency, even if it means disclosing some sensitive information
      - ☐ Balanced approach, protecting critical data while providing sufficient explanations
      - ☐ Strict confidentiality, limiting explanations to preserve data privacy
   2. Simplicity vs. Completeness
      - ☐ Prioritize simple, easily understandable explanations
      - ☐ Strike a balance between simplicity and completeness
      - ☐ Provide comprehensive, detailed explanations, even if they are more complex
   3. Global vs. Local Explanations
      - ☐ Focus on global explanations that describe overall patterns and trends
      - ☐ Balance global and local explanations to address both general insights and specific cases
      - ☐ Prioritize local explanations that provide details on individual predictions or anomalies

We hereby detail a corresponding table that expands on the decision logic for each section of the self-assessment questions:

**Project Details - Decision Logic:**

- If the project involves predictive monitoring or root cause analysis, prioritize XAI techniques like SHAP, LIME, and counterfactual explanations to provide insights into the factors influencing predictions and outcomes. These techniques can help identify the key drivers of process performance and support proactive decision-making (Chapter 2, [163, 164, 379]).

- If the data is unstructured or complex, recommend advanced data preprocessing techniques like natural language processing, pattern mining, and semantic analysis to extract meaningful features and relationships. These techniques can help transform raw data into structured, interpretable formats that enable more accurate and comprehensible explanations (Chapter 2, [150, 278]).
- If the event log is large, consider using distributed computing frameworks and big data technologies to handle the scale and complexity of the data efficiently. Parallel processing and data partitioning strategies can help speed up the explanation generation process and ensure the scalability of the PM-XAI solution (Chapter 2, [300]).
- If specific data attributes like resource or cost are crucial for the analysis, ensure that the selected XAI techniques can effectively incorporate and explain the impact of these factors on process outcomes. Consider using feature importance and sensitivity analysis methods to quantify the relative influence of different attributes on the explanations (Chapter 2, [163, 164]).

**Organizational Context - Decision Logic:**

- If the organization is in a highly regulated industry like healthcare or finance, prioritize compliance with relevant regulations (e.g., GDPR, HIPAA, PCI-DSS) and establish strict data governance and security measures to protect sensitive information. Ensure that the PM-XAI solution includes robust access controls, data encryption, and audit trails to meet regulatory requirements (Chapter 4, [132, 358]).
- If the organization has limited process mining experience, recommend a phased implementation approach, starting with small-scale pilot projects and gradually expanding to more complex use cases. Provide training and change management strategies to build internal capabilities and foster adoption. Use the pilot projects to demonstrate the value and feasibility of PM-XAI and gather lessons learned for future scaling (Chapter 3, [163, 404]).
- If the level of executive support is low, focus on demonstrating the value and ROI of process mining and XAI through targeted case studies and stakeholder engagement. Align the initiative with strategic business objectives and communicate the benefits in terms that resonate with decision-makers. Identify and engage key influencers and sponsors who can champion the PM-XAI initiative and help secure the necessary resources and support (Chapter 3, [339]).
- If the organization has advanced process mining maturity, consider implementing enterprise-wide XAI solutions that integrate with existing tools and workflows. Establish a center of excellence to promote best practices, innovation, and continuous improvement. Leverage the organization's existing PM capabilities and infrastructure to accelerate the adoption and scaling of XAI techniques across different business units and processes (Chapter 3, [399]).

**Explainability Requirements - Decision Logic:**

- If the primary stakeholders are business users or process owners, prioritize descriptive and diagnostic explanations that provide insights into process patterns, deviations, and performance drivers. Use intuitive visualizations and natural language explanations to make the insights accessible and actionable. Focus on explaining the key factors influencing process outcomes and identifying opportunities for improvement (Chapter 3, [163, 164]).

- If the project involves predictive modeling or prescriptive analytics, recommend XAI techniques that provide counterfactual explanations and "what-if" scenario analysis. Enable users to explore the impact of different process configurations and decisions on predicted outcomes. Allow them to simulate alternative scenarios and understand the trade-offs and implications of different choices (Chapter 2, [8, 66, 339]).
- If compliance with regulations like GDPR is a key requirement, ensure that the XAI techniques generate explanations that meet the standards for transparency, fairness, and accountability. Document the explanations and make them available to data subjects upon request. Implement mechanisms for data subjects to challenge or seek human review of automated decisions based on the explanations (Chapter 4, [132, 358]).
- If the process mining insights will be used for decision support or automation, prioritize explanations that build trust and confidence in the recommendations. Provide clear rationales for the suggested actions and allow users to drill down into the underlying data and logic. Enable users to provide feedback and incorporate their domain knowledge to refine the explanations and improve the decision-making process (Chapter 3, [163, 404]).

**Constraints and Priorities - Decision Logic:**

- If limited technical expertise is a constraint, recommend user-friendly, low-code XAI tools that provide pre-built explanation templates and workflows. Offer training and support resources to help users navigate the tools and interpret the results effectively. Consider partnering with external XAI experts or service providers to accelerate the implementation and knowledge transfer (Chapter 3, [143, 404]).
- If scalability and performance are critical, consider deploying the PM-XAI solution on cloud-based infrastructure that can dynamically scale to handle large volumes of data and concurrent users. Optimize the data processing pipeline and explanation generation algorithms to minimize latency. Implement caching and data compression techniques to improve the efficiency and responsiveness of the solution (Chapter 2, [300]).
- If resistance to change is a significant barrier, develop a comprehensive change management plan that addresses the concerns and needs of different stakeholder groups. Engage users early and often, provide clear communication and training, and celebrate quick wins to build momentum and support. Identify and empower change agents who can help drive the adoption and institutionalization of the PM-XAI practices (Chapter 3, [399]).
- If compliance with regulations is a top priority, establish a dedicated governance team to oversee the PM-XAI implementation and ensure adherence to relevant standards and guidelines. If the project requires a balance between transparency and confidentiality, consider using XAI techniques that provide aggregate-level explanations or anonymize sensitive data. Implement access controls and data masking to protect confidential information while still enabling meaningful explanations. Establish clear data sharing and usage agreements with stakeholders to ensure the responsible and secure handling of the explanations (Chapter 4, [115, 380]).

# Appendix D

# List of Publications, Author's Contribution and Copyright Information

This Ph.D dissertation reproduces the contents of the following publications:

- Chapter 4 reproduces the contents of Luca Nannini, Agathe Balayn, Adam Leon Smith *Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK*, *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency - FAccT '23*, 1198-1212, 2023, Chicago, IL, USA. (Available online: June 2023). ISBN: 979-8-4007-0192-4. doi: 10.1109/MIS.2024.3383155. Quality Factor: N/A. Reproduction rights: This article has been published Open Access.

- **Author's contributions**:

  **Luca Nannini:** Conceptualization, Resources, Writing — Review & Editing, Visualization, Project administration, Funding acquisition, Investigation.

  **Agathe Balayn:** Writing — Review & Editing, Visualization, Investigation.

  **Adam Leon Smith:** Writing — Review & Editing, Visualization, Investigation.

  **Reproduction authorization:** ACM allows to reuse accepted papers in a Ph.D dissertation without express permission.

  See https://www.acm.org/publication-rights-and-licensing-policy and Figure D.1.

## PERMISSIONS

ACM grants gratis permission for individual digital or hard copies made without fee for use in academic classrooms and for use by individuals in personal research and study. Further reproduction or distribution requires explicit permission and possibly a fee.

ACM is now a signatory of the STM Permission Guidelines Initiative, which supports an approach to research based on common decency, respect, fairness and mutual trust. These Guidelines are built to allow Signatory STM Publishers to use limited amounts of material in other original published works without charge, and with a minimum of effort needed for permissions clearance. ACM joined the initiative in 2022 to lower the burden on authors to obtain third party permissions when authoring works for ACM and third party publishers.

All copies should carry the original citation, the appropriate copyright and notice of permission on the first page or initial screen of the document. (See §2.2 Copyright Notice.)

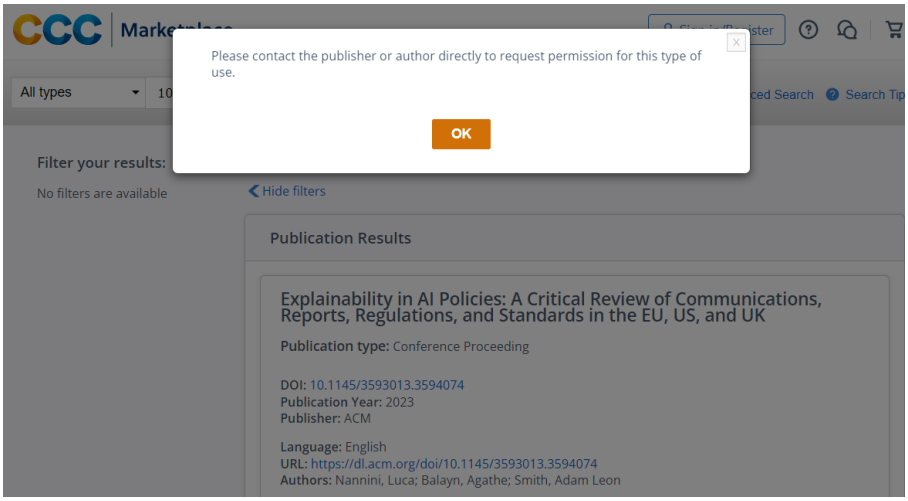


**Figure D.1:** ACM Permission

- Chapter 4 reproduces the contents of Luca Nannini, Jose Maria Alonso-Moral, Alejandro Catala, Manuel Lama, and Senén Barro, *Operationalizing Explainable AI in the EU Regulatory Ecosystem*, *IEEE Intelligent Systems*, 1-13, 2024 (Available online: April 2024). ISSN: 1541-1672. doi: 10.1109/MIS.2024.3383155. Journal Impact Factor: 5.6 (JCR 2023), 2.11 (SJR 2023). Journal ranked in JCR 2023, in Computer science, Artificial Intelligence, (Q1, 41/197); in Engineering, Electrical & Electronic (Q1, 53/352).

- **Author's contributions**:

  **Luca Nannini:** Methodology, Formal Analysis, Investigation, Data Curation, Writing — Original Draft.

**Jose Maria Alonso-Moral:** Conceptualization, Resources, Writing — Review & Editing, Visualization, Supervision, Project administration, Funding acquisition, Investigation

**Alejandro Catala:** Conceptualization, Resources, Writing — Review & Editing, Visualization, Supervision, Project administration, Funding acquisition, Investigation.

**Manuel Lama:** Conceptualization, Resources, Writing — Review & Editing, Supervision, Project administration, Funding acquisition, Investigation.

**Senén Barro:** Conceptualization, Resources, Writing — Review & Editing, Supervision, Project administration, Funding acquisition, Investigation.

**Reproduction authorization:** IEEE allows to reuse accepted papers in a Ph.D dissertation without express permission.

See https://journals.ieeeauthorcenter. ieee.org/choose-a-publishing-agreement/avoid-infringemen t-upon-ieee-copyright and Figure D.2.

> In reference to IEEE copyrighted material which is used with permision in this thesis, the IEEE does not endorse any of Universidade de Santiago's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works fr resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

**Thesis / Dissertation Reuse**

**The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:**

*Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:*

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

*Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:*

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

**Figure D.2:** IEEE Permission

- Chapter 5 reproduces the contents of Luca Nannini, Marta Marchiori Manerba, and Isacco Beretta, *Mapping the Landscape of Ethical Considerations in Explainable AI Research*, Springer-Verlag Ethics and Information Technology, 1-22, 2024 (Available online: June 2024). ISSN: 1388-1957. doi: 10.1007/s10676-024-09773-7. Journal Impact Factor: 3.6 (JCR 2023). Journal ranked in JCR 2023, in Ethics (Q1, 8/57); in Information Science & Library Science; in Philosophy (N/A).

- **Author's contributions**:

  **Luca Nannini:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing — Original Draft.

  **Marta Marchiori Manerba:** Resources, Writing — Review & Editing, Visualization, Funding acquisition, Investigation.

  **Isacco Beretta:** Resources, Writing — Review & Editing, Visualization, Funding acquisition, Investigation.

**Reproduction authorization:** Springer Nature allows to reuse accepted papers in a Ph.D dissertation without express permission.

See https://www.springer.com/gp/rights-permissions/obtaining-permissions/882 and Figure D.3.

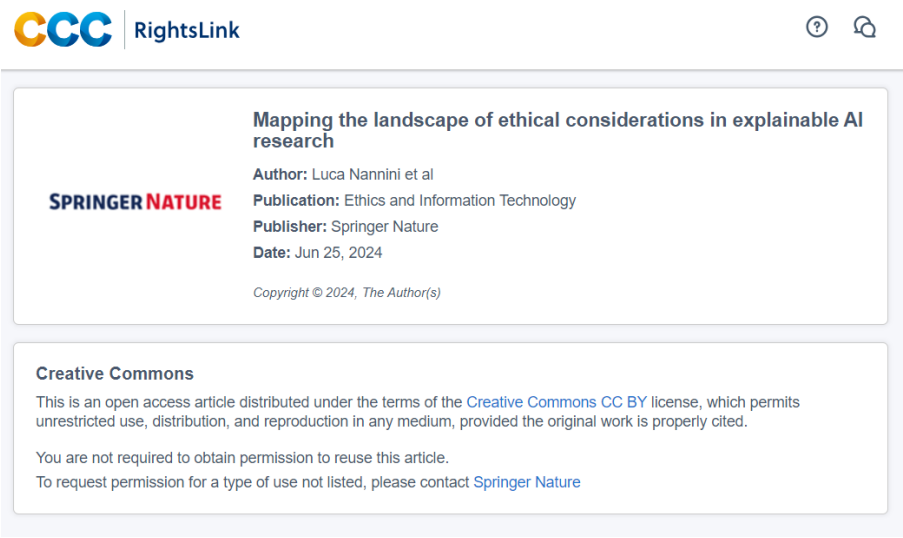


**Figure D.3:** Springer Nature Permission

# Bibliography


[1] European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. Draft standardisation request to the european standardisation organisations in support of safe and trustworthy artificial intelligence, 2022.

[2] 116th Congress of the United States of America. National artificial intelligence initiative act of 2020, 2020.

[3] 117th Congress of the United States of America. Algorithmic accountability act of 2022, h.r.6580, 2022.

[4] Rafael Accorsi and Julian Lebherz. A practitioner's view on process mining adoption, event log engineering and data challenges. In van der Aalst and Carmona [469], pages 212–240.

[5] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[6] Jan Niklas Adams, Gyunam Park, Sergej Levich, Daniel Schuster, and Wil M. P. van der Aalst. A framework for extracting and encoding features from object-centric event data. In Javier Troya, Brahim Medjahed, Mario Piattini, Lina Yao, Pablo Fernández, and Antonio Ruiz-Cortés, editors, *Service-Oriented Computing - 20th International Conference, ICSOC 2022, Seville, Spain, November 29 - December 2, 2022, Proceedings*, volume 13740 of *Lecture Notes in Computer Science*, pages 36–53. Springer, 2022.

[7] Jan Niklas Adams, Sebastiaan J. van Zelst, Lara Quack, Kathrin Hausmann, Wil M. P. van der Aalst, and Thomas Rose. A framework for explainable

concept drift detection in process mining. In Polyvyanyy et al. [388], pages 400–416.

[8] Jan Niklas Adams, Sebastiaan J. van Zelst, Thomas Rose, and Wil M. P. van der Aalst. Explainable concept drift in process mining. *Inf. Syst.*, 114:102177, 2023.

[9] Robert Merrihew Adams. Motive utilitarianism. *The Journal of Philosophy*, 73(14):467–481, 1976.

[10] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9525–9536, 2018.

[11] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[12] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[13] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowl. Inf. Syst.*, 54(1):95–122, 2018.

[14] Chirag Agarwal, Marinka Zitnik, and Himabindu Lakkaraju. Probing GNN explainers: A rigorous theoretical and empirical analysis of GNN explanation methods. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 8969–8996. PMLR, 2022.

[15] Agencia Estatal - Boletín Oficial del Estado. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, 2018.

[16] Philip E Agre et al. Lessons learned in trying to reform ai. *Social science, technical systems, and cooperative work: Beyond the Great Divide*, 131, 1997.

[17] Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where. *IEEE Trans. Ind. Informatics*, 18(8):5031–5042, 2022.

[18] Mhairi Aitken, David Leslie, Florian Ostmann, Jacob Pratt, Helen Margetts, and Cosmina Dorobantu. Common regulatory capacity for ai. Technical report, The Alan Turing Institute, 2022.

[19] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 161–170. PMLR, 2019.

[20] Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. Characterizing the risk of fairwashing. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 14822–14834, 2021.

[21] Hanan Alkhammash, Artem Polyvyanyy, Alistair Moffat, and Luciano García-Bañuelos. Entropic relevance: A mechanism for measuring stochastic process models discovered from event data. *Inf. Syst.*, 107:101922, 2022.

[22] José Maria Alonso, J. Toja-Alamancos, and Alberto Bugarín. Experimental study on generating multi-modal explanations of black-box classifiers in terms of gray-box classifiers. In *FUZZ-IEEE*, pages 1–8. IEEE, 2020.

[23] Ricardo S. Alonso. Deep symbolic learning and semantics for an explainable and ethical artificial intelligence. In Paulo Novais, Gianni Viardo Vercelli,

Josep Lluís Larriba-Pey, Francisco Herrera, and Pablo Chamoso, editors, *Ambient Intelligence - Software and Applications - 11th International Symposium on Ambient Intelligence, ISAmI 2020, L'Aquila, Italy, October 7 - 9, 2020*, volume 1239 of *Advances in Intelligent Systems and Computing*, pages 272–278. Springer, 2020.

[24] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7786–7795, 2018.

[25] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics Decis. Mak.*, 20(1):310, 2020.

[26] Guy Amit, Fabiana Fournier, Shlomit Gur, and Lior Limonad. Model-informed LIME extension for business process explainability. In Giuseppe De Giacomo, Antonella Guzzo, Marco Montali, Lior Limonad, Fabiana Fournier, and Tagatha Chakraborti, editors, *Proceedings of the Workshop on Process Management in the AI Era (PMAI 2022) co-located with 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence (IJCAI-ECAI 2022), Wien, Austria, July 23, 2022*, volume 3310 of *CEUR Workshop Proceedings*, pages 1–12. CEUR-WS.org, 2022.

[27] Guy Amit, Fabiana Fournier, Lior Limonad, and Inna Skarbovsky. Situation-aware explainability for business processes enabled by complex events. In Cristina Cabanillas, Niels Frederik Garmann-Johnsen, and Agnes Koschmider, editors, *Business Process Management Workshops - BPM 2022 International Workshops, M¨unster, Germany, September 11-16, 2022, Revised Selected Papers*, volume 460 of *Lecture Notes in Business Information Processing*, pages 45–57. Springer, 2022.

[28] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.*, 20(3):973–989, 2018.

[29] Gertrude Elizabeth Margaret Anscombe. Modern moral philosophy. *Philosophy*, 33(124):1–19, 1958.

[30] Pedro Antunes, José A. Pino, and Mary Tate. Method for eliciting and analyzing business processes based on storytelling theory. In Tung Bui, editor, *52nd Hawaii International Conference on System Sciences, HICSS 2019, Grand Wailea, Maui, Hawaii, USA, January 8-11, 2019*, pages 1–10. ScholarSpace, 2019.

[31] Daniel Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 06 2020.

[32] Achilles A Armenakis and Stanley G Harris. Crafting a change message to create transformational readiness. *Journal of organizational change management*, 15(2):169–183, 2002.

[33] Chetan Arora, Mehrdad Sabetzadeh, Lionel C. Briand, and Frank Zimmer. Extracting domain models from natural-language requirements: approach and industrial evaluation. In Benoit Baudry and Benoît Combemale, editors, *Proceedings of the ACM/IEEE 19th International Conference on Model Driven Engineering Languages and Systems, Saint-Malo, France, October 2-7, 2016*, pages 250–260. ACM, 2016.

[34] Siddhant Arora, Danish Pruthi, Norman M. Sadeh, William W. Cohen, Zachary C. Lipton, and Graham Neubig. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5277–5285. AAAI Press, 2022.

[35] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115, 2020.

[36] XAI Explainable Artificial Intelligence IEEE Computer Society (IEEE C/AIS-C/XAI) Artificial Intelligence Standards Committee. Ieee p2894 - guide for an architectural framework for explainable artificial intelligence. `https://standards.ieee.org/ieee/2894/10284/`, 2020.

[37] Adriano Augusto, Raffaele Conforti, Marlon Dumas, Marcello La Rosa, and Artem Polyvyanyy. Split miner: automated discovery of accurate and simple business process models from event logs. *Knowl. Inf. Syst.*, 59(2):251–284, 2019.

[38] Jacqui Ayling and Adriane Chapman. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics*, September 2021.

[39] Amgad Badewi. The impact of project management (pm) and benefits management (bm) practices on project success: Towards developing a project benefits governance framework. *International Journal of Project Management*, 34(4):761–778, 2016.

[40] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[41] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1194–1206. ACM, 2022.

[42] Nagadivya Balasubramaniam, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, and Sari Kujala. Transparency and explainability of AI systems: From ethical guidelines to requirements. *Inf. Softw. Technol.*, 159:107197, 2023.

[43] Hubert Baniecki, Wojciech Kretowicz, and Przemyslaw Biecek. Fooling partial dependence via data poisoning. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022,*

*Proceedings, Part III*, volume 13715 of *Lecture Notes in Computer Science*, pages 121–136. Springer, 2022.

[44] László Bántay and János Abonyi. Frequent pattern mining-based log file partition for process mining. *Eng. Appl. Artif. Intell.*, 123(Part A):106221, 2023.

[45] Luciana Barbieri, Edmundo R. M. Madeira, Kleber Stroeh, and Wil M. P. van der Aalst. A natural language querying interface for process mining. *J. Intell. Inf. Syst.*, 61(1):113–142, 2023.

[46] Anton Batliner, Simone Hantke, and Björn W. Schuller. Ethics and good practice in computational paralinguistics. *IEEE Trans. Affect. Comput.*, 13(3):1236–1253, 2022.

[47] Kevin Baum, Susanne Mantel, Timo Speith, and Eva Schmidt. From responsibility to reason-giving explainable artificial intelligence. *Philosophy and Technology*, 35(1):1–30, 2022.

[48] Tom L Beauchamp and James F Childress. *Principles of biomedical ethics*. Oxford University Press, USA, 2001.

[49] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. Dover Publications, New York, 1780.

[50] Christoph Benzmüller and Bertram Lomfeld. Reasonable machines: A research manifesto. In Ute Schmid, Franziska Klügl, and Diedrich Wolter, editors, *KI 2020: Advances in Artificial Intelligence - 43rd German Conference on AI, Bamberg, Germany, September 21-25, 2020, Proceedings*, volume 12325 of *Lecture Notes in Computer Science*, pages 251–258, Germany, 2020. Springer.

[51] Vered Bernstein and Pnina Soffer. How does it look? exploring meaningful layout features of process models. In Anne Persson and Janis Stirna, editors, *Advanced Information Systems Engineering Workshops*, Lecture Notes in Business Information Processing, pages 81–86. Springer International Publishing, 2015.

[52] Elettra Bietti. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 210–219. ACM, 2020.

[53] Kristof Böhmer and Stefanie Rinderle-Ma. Logo: Combining local and global techniques for predictive business process monitoring. In Schahram Dustdar, Eric Yu, Camille Salinesi, Dominique Rieu, and Vik Pant, editors, *Advanced Information Systems Engineering - 32nd International Conference, CAiSE 2020, Grenoble, France, June 8-12, 2020, Proceedings*, volume 12127 of *Lecture Notes in Computer Science*, pages 283–298. Springer, 2020.

[54] Andrew Booth, Anthea. Sutton, and Diana. Papaioannou. *Systematic approaches to a successful literature review*. SAGE, Los Angeles, second edition. edition, 2016.

[55] Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 891–905, New York, NY, USA, 2022. Association for Computing Machinery.

[56] Alessio Bottrighi, Luca Canensi, Giorgio Leonardi, Stefania Montani, and Paolo Terenziani. Interactive mining and retrieval from process traces. *Expert Syst. Appl.*, 110:62–79, 2018.

[57] Joshua L. M. Brand and Luca Nannini. Does explainable AI have moral value? *CoRR*, abs/2311.14687, 2023.

[58] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:77–101, 2006.

[59] Alejandra Bringas Colmenarejo, Luca Nannini, Alisa Rieger, Kristen M. Scott, Xuan Zhao, Gourab K Patro, Gjergji Kasneci, and Katharina Kinder-Kurlanda. Fairness in agreement with european values: An interdisciplinary perspective on ai regulation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 107–118, New York, NY, USA, 2022. Association for Computing Machinery.

[60] David Broniatowski. Psychological foundations of explainability and interpretability in artificial intelligence. Technical report, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, 2021-04-12 04:04:00 2021.

[61] Davis Brown and Henry Kvinge. Making corgis important for honeycomb classification: Adversarial attacks on concept-based explainability tools. In

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 620–627, 2023.

[62] Jens Brunk, Matthias Stierle, Leon Papke, Kate Revoredo, Martin Matzner, and J¨org Becker. Cause vs. effect in context-sensitive prediction of business process instances. *Inf. Syst.*, 95:101635, 2021.

[63] Wasja Brunotte, Larissa Chazette, Verena Klös, and Timo Speith. Quo vadis, explainability? - A research roadmap for explainability engineering. In Vincenzo Gervasi and Andreas Vogelsang, editors, *Requirements Engineering: Foundation for Software Quality - 28th International Working Conference, REFSQ 2022, Birmingham, UK, March 21-24, 2022, Proceedings*, volume 13216 of *Lecture Notes in Computer Science*, pages 26–32. Springer, 2022.

[64] Tobias Budig, Selina Herrmann, and Alexander Dietz. Trade-offs between privacy-preserving and explainable machine learning in healthcare. In *Seminar Paper, Inst. Appl. Informat. Formal Description Methods (AIFB), KIT Dept. Econom. Manage., Karlsruhe, Germany*, 2020.

[65] Joos C. A. M. Buijs, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity. *Int. J. Cooperative Inf. Syst.*, 23(1), 2014.

[66] Andrei Buliga, Chiara Di Francescomarino, Chiara Ghidini, and Fabrizio Maria Maggi. Counterfactuals and ways to build them: Evaluating approaches predictive process monitoring. In *Advanced Information Systems Engineering: 35th International Conference, CAiSE 2023, Zaragoza, Spain, June 12–16, 2023, Proceedings*, page 558–574, Berlin, Heidelberg, 2023. Springer-Verlag.

[67] Andrea Burattin, Artem Polyvyanyy, and Barbara Weber, editors. *4th International Conference on Process Mining, ICPM 2022, Bolzano, Italy, October 23-28, 2022*. IEEE, 2022.

[68] Kiran Busch, Alexander Rochlitzer, Diana Sola, and Henrik Leopold. Just tell me: Prompt engineering in business process management. In Han van der Aa, Dominik Bork, Henderik A. Proper, and Rainer Schmidt, editors, *Enterprise, Business-Process and Information Systems Modeling - 24th International Conference, BPMDS 2023, and 28th International Conference, EMMSAD 2023, Zaragoza, Spain, June 12-13, 2023, Proceedings*, volume 479 of *Lecture Notes in Business Information Processing*, pages 3–11. Springer, 2023.

[69] Federico Cabitza, Andrea Campagner, Gianclaudio Malgieri, Chiara Natali, David Schneeberger, Karl Stöger, and Andreas Holzinger. Quod erat demonstrandum? - towards a typology of the concept of explanation for the design of explainable AI. *Expert Syst. Appl.*, 213(Part):118888, 2023.

[70] Roberta Calegari, Andrea Omicini, and Giovanni Sartor. Argumentation and logic programming for explainable and ethical AI. In Cataldo Musto, Daniele Magazzeni, Salvatore Ruggieri, and Giovanni Semeraro, editors, *Proceedings of the Italian Workshop on Explainable Artificial Intelligence co-located with 19th International Conference of the Italian Association for Artificial Intelligence, XAI.it@AIxIA 2020, Online Event, November 25-26, 2020*, volume 2742 of *CEUR Workshop Proceedings*, pages 55–68. CEUR-WS.org, 2020.

[71] Josep Carmona, Jordi Cortadella, and Michael Kishinevsky. A region-based algorithm for discovering petri nets from event logs. In Marlon Dumas, Manfred Reichert, and Ming-Chien Shan, editors, *Business Process Management, 6th International Conference, BPM 2008, Milan, Italy, September 2-4, 2008. Proceedings*, volume 5240 of *Lecture Notes in Computer Science*, pages 358–373. Springer, 2008.

[72] Josep Carmona, Boudewijn F. van Dongen, and Matthias Weidlich. Conformance checking: Foundations, milestones and challenges. In van der Aalst and Carmona [469], pages 155–190.

[73] D.V. Carvalho, E.M. Pereira, and J.S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)*, 8(8), 2019. Publisher: MDPI AG.

[74] CEN-CENELEC. Focus group report - road map on artificial intelligence (ai), 2020. Accessed on: 2023-01-30.

[75] CEN-CENELEC. Joint technical committee 21 (jtc 21) 'artificial intelligence', 2021. Accessed on: 2023-01-30.

[76] Central Digital & Data Office & Centre for Data Ethics & Innovation. Algorithmic transparency recording standard - guidance, 2021.

[77] Kathy Charmaz. Constructionism and the grounded theory method. *Handbook of constructionist research*, 1(1):397–412, 2008.

[78] Larissa Chazette, Oliver Karras, and Kurt Schneider. Do end-users want explanations? analyzing the role of explainability as an emerging aspect of non-functional requirements. In Daniela E. Damian, Anna Perini, and Seok-Won Lee, editors, *27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea (South), September 23-27, 2019*, pages 223–233, Jeju Island, Korea (South), 2019. IEEE.

[79] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Dy and Krause [114], pages 882–891.

[80] Sudhanshu G. R. Chouhan, Anna Wilbik, and Remco M. Dijkman. Explanation of anomalies in business process event logs with linguistic summaries. In *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padua, Italy, July 18-23, 2022*, pages 1–7. IEEE, 2022.

[81] Claudio Di Ciccio, Chiara Di Francescomarino, and Pnina Soffer, editors. *3rd International Conference on Process Mining, ICPM 2021, Eindhoven, The Netherlands, October 31 - Nov. 4, 2021*. IEEE, 2021.

[82] Herbert H. Clark and Susan Brennan. Grounding in communication. In *Perspectives on socially shared cognition*, 1991.

[83] Jennifer Cobbe, Michael Veale, and Jatinder Singh. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 1186–1197. ACM, 2023.

[84] William W. Cohen. Fast effective rule induction. In Armand Prieditis and Stuart Russell, editors, *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pages 115–123. Morgan Kaufmann, 1995.

[85] European Commission. Communication Artificial Intelligence for Europe: Shaping Europe's digital future, April 2018.

[86] European Commission. White paper on artificial intelligence: A european approach to excellence and trust. *Brussels, 1st edn. European Commission, Brussels*, 2020.

[87] European Commission. Communication on fostering a european approach to artificial intelligence, 2021.

[88] European Commission. Coordinated plan on artificial intelligence 2021 review, 2021.

[89] European Commission. Rolling plan for ict standardisation 2022, 2022.

[90] European Commission, Joint Research Centre, S Nativi, and S De Nigris. Ai watch, ai standardisation landscape state of play and link to the ec proposal for an ai regulatory framework. Technical report, Publications Office of the European Union, 2021.

[91] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. A historical perspective of explainable artificial intelligence. *WIREs Data Mining Knowl. Discov.*, 11(1), 2021.

[92] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[93] Council of the European Union. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts - general approach. https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf, 11 2022. LIMITE TELECOM 472 JAI 1494 COPEN 396 CYBER 374 DATAPROTECT 320 EJUSTICE 89 COSI 293 IXIM 267 ENFOPOL 569 RELEX 1556 MI 843 COMPET 918 CODEC 1773. From: Permanent Representatives Committee (Part 1). Interinstitutional File: 2021/0106(COD).

[94] Alfredo Cuzzocrea, Francesco Folino, Massimo Guarascio, and Luigi Pontieri. Experimenting and assessing a probabilistic business process deviance mining framework based on ensemble learning. In Slimane Hammoudi, Michal Smialek, Olivier Camp, and Joaquim Filipe, editors, *Enterprise Information Systems - 19th International Conference, ICEIS 2017, Porto, Portugal, April 26-29, 2017, Revised Selected Papers*, volume 321 of *Lecture Notes in Business Information Processing*, pages 96–124. Springer, 2017.

[95] Jessica Dai, Sohini Upadhyay, Ulrich Aïvodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In Vincent Conitzer, John Tasioulas, Matthias

Scheutz, Ryan Calo, Martina Mara, and Annette Zimmermann, editors, *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*, pages 203–214. ACM, 2022.

[96] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525, 2021.

[97] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2):101666, 2022.

[98] Giovanni De Gregorio and Simona Demkova. The Constitutional Right to an Effective Remedy in the Digital Age: A Perspective from Europe. In Ch. van Oirsouw, J. de Poorter, I. Leijten, G. van der Schyff, M. Stremler, and M. de Visser, editors, *European Yearbook of Constitutional Law*. Forthcoming, January 2024. Available at SSRN.

[99] Massimiliano de Leoni. Foundations of process enhancement. In van der Aalst and Carmona [469], pages 243–273.

[100] Hugo De Oliveira, Martin Prodel, Ludovic Lamarsalle, Vincent Augusto, and Xiaolan Xie. Explaining predictive factors in patient pathways using autoencoders. *PLOS ONE*, 17:e0277135, 11 2022.

[101] Pavlos Delias, Vassilios Zoumpoulidis, and Ioannis Kazanidis. Visualizing and exploring event databases: a methodology to benefit from process analytics. *Oper. Res.*, 19(4):887–908, 2019.

[102] Luis Delicado, Josep Sànchez-Ferreres, Josep Carmona, and Lluís Padró. NLP4BPM - natural language processing tools for business process management. In Robert Clarisó, Henrik Leopold, Jan Mendling, Wil M. P. van der Aalst, Akhil Kumar, Brian T. Pentland, and Mathias Weske, editors, *Proceedings of the BPM Demo Track and BPM Dissertation Award co-located with 15th International Conference on Business Process Modeling (BPM 2017), Barcelona, Spain, September 13, 2017*, volume 1920 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2017.

[103] Jacques Derrida. *Dissemination*. Bloomsbury Publishing, 2016.

[104] Patricia G Devine. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5, 1989.

[105] Jacob Dexe, Ulrik Franke, Anneli Avatare Nöu, and Alexander Rad. Towards increased transparency with value sensitive design. In Helmut Degen and Lauren Reinerman-Jones, editors, *Artificial Intelligence in HCI - First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings*, volume 12217 of *Lecture Notes in Computer Science*, pages 3–15, Denmark, 2020. Springer.

[106] M. Diahame, M. Diop, and M.S. Camara. Prediction of the failure of a business process event using event log and deep learning. *Lecture Notes in Networks and Systems*, 596 LNNS:266–276, 2023.

[107] Ahmet Dikici, Oktay Türetken, and Onur Demirörs. Factors influencing the understandability of process models: A systematic literature review. *Inf. Softw. Technol.*, 93:112–129, 2018.

[108] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In Huáscar Espinoza, José Hernández-Orallo, Xin Cynthia Chen, Seán S. ÓhÉigeartaigh, Xiaowei Huang, Mauricio Castillo-Effen, Richard Mallah, and John A. McDermid, editors, *Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020, New York City, NY, USA, February 7, 2020*, volume 2560 of *CEUR Workshop Proceedings*, pages 63–73. CEUR-WS.org, 2020.

[109] Lex Donaldson. *The contingency theory of organizations*. Sage, 2001.

[110] Vasisht Duddu and Antoine Boutet. Inferring sensitive attributes from model explanations. In Mohammad Al Hasan and Li Xiong, editors, *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 416–425. ACM, 2022.

[111] Marlon Dumas, Marcello La Rosa, Jan Mendling, Raul Mäesalu, Hajo A. Reijers, and Nataliia Semenenko. Understanding business process models: The costs and benefits of structuredness. In Jolita Ralyté, Xavier Franch, Sjaak

Brinkkemper, and Stanislaw Wrycza, editors, *Advanced Information Systems Engineering - 24th International Conference, CAiSE 2012, Gdansk, Poland, June 25-29, 2012. Proceedings*, volume 7328 of *Lecture Notes in Computer Science*, pages 31–46. Springer, 2012.

[112] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), jan 2023.

[113] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

[114] Jennifer G. Dy and Andreas Krause, editors. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018.

[115] Martin Ebers. Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s), August 2021.

[116] Rebekah Eden, Rehan Syed, Sander J. J. Leemans, and Joos A. C. M. Buijs. A case study of inconsistency in process mining use: Implications for the theory of effective use. In Polyvyanyy et al. [388], pages 363–379.

[117] Lilian Edwards and Michael Veale. Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE Security & Privacy*, 16(3):46–54, May 2018. Conference Name: IEEE Security & Privacy.

[118] Julia Eggers, Andreas Hein, Markus Böhm, and Helmut Krcmar. No longer out of sight, no longer out of mind? how organizations engage with process mining-induced transparency to achieve increased process awareness. *Bus Inf Syst Eng*, 63(5):491–510, 2021.

[119] Mansoureh Yari Eili and Jalal Rezaeenour. A survey on recommendation in process mining. *Concurr. Comput. Pract. Exp.*, 34(26), 2022.

[120] Ray Eitel-Porter. Beyond the promise: implementing ethical AI. *AI Ethics*, 1(1):73–80, February 2021.

[121] Magy Seif El-Nasr and Erica Kleinman. Data-driven game development: Ethical considerations. In Georgios N. Yannakakis, Antonios Liapis, Penny Kyburz, Vanessa Volz, Foaad Khosmood, and Phil Lopes, editors, *FDG '20: International Conference on the Foundations of Digital Games, Bugibba, Malta, September 15-18, 2020*, pages 64:1–64:10, Malta, 2020. ACM.

[122] Ghada Elkhawaga, Omar M Elzeki, Mervat Abu-Elkheir, and Manfred Reichert. Why should i trust your explanation? an evaluation approach for xai methods applied to predictive process monitoring results. *IEEE Transactions on Artificial Intelligence*, 1(01):1–15, jan 2024.

[123] Daniel C. Elton. Common Pitfalls When Explaining AI and Why Mechanistic Explanation Is a Hard Problem. In Xin-She Yang, Simon Sherratt, Nilanjan Dey, and Amit Joshi, editors, *Proceedings of Sixth International Congress on Information and Communication Technology*, volume 235, pages 401–408. Springer Singapore, Singapore, 2022. Series Title: Lecture Notes in Networks and Systems.

[124] Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 01 2009.

[125] European Telecommunications Standards Institute (ETSI). ETSI securing artificial intelligence (sai) committee, 2019.

[126] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-COM-Proposal-21-April-21.pdf, 4 2021. 2021/0106 (COD), SEC(2021) 167 final, SWD(2021) 84 final, SWD(2021) 85 final.

[127] European Parliament. Artificial intelligence act amendments adopted by the european parliament on 14 june 2023 on the proposal for a regulation of the european parliament and of the council on laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Texts Adopted P9_TA(2023)0236, 6 2023. COM(2021)0206 - C9-0146/2021 - 2021/0106(COD). Ordinary legislative procedure: first reading.

The matter was referred back for interinstitutional negotiations to the committee responsible, pursuant to Rule 59(4), fourth subparagraph (A9-0188/2023).

[128] European Parliament. Artificial intelligence act: deal on comprehensive rules for trustworthy ai. Press Release, 12 2023. IMCO LIBE.

[129] European Parliament. Artificial intelligence act: MEPs adopt landmark law. Press Release, 3 2024. PLENARY SESSION IMCO LIBE.

[130] European Parliament. Corrigendum to the position of the european parliament adopted at first reading on 13 march 2024 with a view to the adoption of regulation (eu) 2024/... of the european parliament and of the council laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act). Plenary sitting cor01, 4 2024. P9_TA(2024)0138, COM(2021)0206 - C9-0146/2021 - 2021/0106(COD).

[131] European Parliament and Council of the European Union. Regulation (eu) 2024/. . . of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), June 2024. LEX 2363, PE-CONS 24/1/24 REV 1.

[132] European Parliament and Council of the 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 L 119, 4.5.2016, p. 1-88, 2016. Text with EEA relevance. In force, with current consolidated version as of 4/5/2016.

[133] European Parliament and Council of the European Union. Proposal for a directive of the european parliament and of the council on liability for defective products, 1 2022. Committee on the Internal Market and Consumer Protection, Committee on Legal Affairs. PE758.731v01-00.

[134] European Parliament Committee on Legal Affairs. Opinion of the committee on legal affairs for the committee on the internal market and consumer protection and the committee on civil liberties, justice and home affairs on the proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://artificialintelligenceact.eu/wp-content/uploads/2022/09/AIA-JURI-Rule-57-Opinion-Adopted-12-September.pdf, 9 2022. Rapporteur for opinion: Axel Voss. COM(2021)0206 - C9-0146/2021 - 2021/0106(COD).

[135] Experiential Networked Intelligence (ENI) European Telecommunications Standards Institute (ETSI). Etsi gs eni 005 v2.1.1 - experiential networked intelligence (eni), system architecture, 2021.

[136] Securing Artificial Intelligence (SAI) European Telecommunications Standards Institute (ETSI). Dgr/sai-007' work item: Explicability and transparency of ai processing, 2023.

[137] Dirk Fahland, Fabiana Fournier, Lior Limonad, Inna Skarbovsky, and Ava J. E. Swevels. Why are my pizzas late? In Fabiana Fournier, Lior Limonad, Massimiliano de Leoni, Antonella Guzzo, Andrea Marrella, and Pnina Soffer, editors, *Proceedings of the 2nd International Workshop on Process Management in the AI Era (PMAI 2023) co-located with 31st International Joint Conference on Artificial Intelligence (IJCAI 2023), Macao, S.A.R, August 19, 2023*, volume 3569 of *CEUR Workshop Proceedings*, pages 25–28. CEUR-WS.org, 2023.

[138] Dirk Fahland, Chiara Ghidini, J¨org Becker, and Marlon Dumas, editors. *Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings*, volume 392 of *Lecture Notes in Business Information Processing*. Springer, 2020.

[139] Zoe Falomir and Vicent Costa. On the rationality of explanations in classification algorithms. In Mateu Villaret, Teresa Alsinet, Cèsar Fernández, and Aïda Valls, editors, *Artificial Intelligence Research and Development - Proceedings of the 23rd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2021, Virtual Event, 20-22 October, 2021*, volume 339 of *Frontiers in Artificial Intelligence and Applications*, pages 445–454. IOS Press, 2021.

[140] Bettina Fazzinga, Sergio Flesca, Filippo Furfaro, and Luigi Pontieri. Process mining meets argumentation: Explainable interpretations of low-level event logs via abstract argumentation. *Inf. Syst.*, 107:101987, 2022.

[141] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Sci. Eng. Ethics*, 26(6):3333–3361, 2020.

[142] Myriel Fichtner and Stefan Jablonski. A specification of how to extract relevant process details to improve process models. In Joaquim Filipe, Michal Smialek, Alexander Brodsky, and Slimane Hammoudi, editors, *Enterprise Information Systems - 24th International Conference, ICEIS 2022, Virtual Event, April 25-27, 2022, Revised Selected Papers*, volume 487 of *Lecture Notes in Business Information Processing*, pages 391–414. Springer, 2022.

[143] Myriel Fichtner, Stefan Sch¨onig, and Stefan Jablonski. How LIME explanation models can be used to extend business process models by relevant process details. In Joaquim Filipe, Michal Smialek, Alexander Brodsky, and Slimane Hammoudi, editors, *Proceedings of the 24th International Conference on Enterprise Information Systems, ICEIS 2022, Online Streaming, April 25-27, 2022, Volume 2*, pages 527–534. SCITEPRESS, 2022.

[144] Kathrin Figl. Comprehension of procedural visual business process models. *Bus Inf Syst Eng*, 59(1):41–67, 2017.

[145] Kathrin Figl and Jan Recker. Exploring cognitive style and task-specific preferences for process representations. *Requirements Engineering*, 21(1):63–85, 2016.

[146] Baruch Fischhoff, Sarah Lichtenstein, Paul Slovic, Stephen L Derby, and Ralph L Keeney. Defining risk. *Policy Sciences*, 17(2):123–139, 1984.

[147] Will Fleisher. Understanding, idealization, and explainable ai. *Episteme*, 19(4):534–560, 2022.

[148] Luciano Floridi. Distributed morality in an information society. *Science and engineering ethics*, 19:727–743, 2013.

[149] Luciano Floridi. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions*

*of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160112, 2016.

[150] Yago Fontenla-Seco, Manuel Lama, and Alberto Bugarín. Process-to-text: A framework for the quantitative description of processes in natural language. In Fredrik Heintz, Michela Milano, and Barry O'Sullivan, editors, *Trustworthy AI - Integrating Learning, Optimization and Reasoning - First International Workshop, TAILOR 2020, Virtual Event, September 4-5, 2020, Revised Selected Papers*, volume 12641 of *Lecture Notes in Computer Science*, pages 212–219. Springer, 2020.

[151] Philippa Foot. *Virtues and vices and other essays in moral philosophy*. University of California Press, 1978.

[152] Department for Digital, Culture, and Media & Sport of the United Kingdom. National data strategy, 2019.

[153] Department for Digital, Culture, Media & Sport, Department for Business, and Energy & Industrial Strategy of the United Kingdom. Ai sector deal - policy paper, 2019.

[154] Department for Digital, Culture, Media & Sport, Department for Business, Energy & Industrial Strategy, and Office for Artificial Intelligence of the United Kingdom. Establishing a pro-innovation approach to regulating ai - policy paper presented to uk parliament, 2022.

[155] Department for Digital, Culture, Media & Sport, Department for Business, Energy & Industrial Strategy, and Office for Artificial Intelligence of the United Kingdom. National ai strategy - ai action plan, 2022.

[156] Standard for XAI eXplainable AI Working Group IEEE Computational Intelligence Society/ Standards Committee (IEEE CIS/SC/XAI WG). Ieee cis/sc/xai wg p2976 - standard for xai – explainable artificial intelligence - for achieving clarity and interoperability of ai systems design, 2024.

[157] Chiara Di Francescomarino, Marlon Dumas, Marco Federici, Chiara Ghidini, Fabrizio Maria Maggi, and Williams Rizzi. Predictive business process monitoring framework with hyperparameter optimization. In Selmin Nurcan, Pnina Soffer, Marko Bajec, and Johann Eder, editors, *Advanced Information Systems Engineering - 28th International Conference, CAiSE 2016, Ljubljana, Slovenia,*

*June 13-17, 2016. Proceedings*, volume 9694 of *Lecture Notes in Computer Science*, pages 361–376. Springer, 2016.

[158] R Edward Freeman, Jeffrey S Harrison, Andrew C Wicks, Bidhan L Parmar, and Simone De Colle. *Stakeholder theory: The state of the art*. Cambridge University Press, 2010.

[159] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[160] Fabian Friedrich, Jan Mendling, and Frank Puhlmann. Process model generation from natural language text. In *International Conference on Advanced Information Systems Engineering*, pages 482–496. Springer, 2011.

[161] Kazuto Fukuchi, Satoshi Hara, and Takanori Maehara. Faking fairness via stealthily biased sampling. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 412–419. AAAI Press, 2020.

[162] Riccardo Galanti, Bernat Coma-Puig, Massimiliano de Leoni, Josep Carmona, and Nicolò Navarin. Explainable predictive process monitoring. In Boudewijn F. van Dongen, Marco Montali, and Moe Thandar Wynn, editors, *2nd International Conference on Process Mining, ICPM 2020, Padua, Italy, October 4-9, 2020*, pages 1–8. IEEE, 2020.

[163] Riccardo Galanti, Massimiliano de Leoni, Merylin Monaro, Nicolò Navarin, Alan Marazzi, Brigida Di Stasi, and Stéphanie Maldera. An explainable decision support system for predictive process analytics. *Eng. Appl. Artif. Intell.*, 120:105904, 2023.

[164] Riccardo Galanti, Massimiliano de Leoni, Nicolò Navarin, and Alan Marazzi. Object-centric process predictive analytics. *Expert Syst. Appl.*, 213(Part):119173, 2023.

[165] Cleiton dos Santos Garcia, Alex Meincheim, Elio Ribeiro Faria Junior, Marcelo Rosano Dallagassa, Denise Maria Vecino Sato, Deborah Ribeiro Carvalho, Eduardo Alves Portela Santos, and Edson Emilio Scalabrin. Process

mining techniques and applications – a systematic mapping study. *Expert Systems with Applications*, 133:260–295, 2019.

[166] Frank W Geels. From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research policy*, 33(6-7):897–920, 2004.

[167] Ilina Georgieva, Claudio Lazo, Tjerk Timan, and Anne Fleur van Veenstra. From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. *AI Ethics*, January 2022.

[168] Anne Gerdes. Dialogical guidelines aided by knowledge acquisition: Enhancing the design of explainable interfaces and algorithmic accuracy. In K. Arai, S. Kapoor, and R. Bhatia, editors, *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1*, volume 1288 of *Advances in Intelligent Systems and Computing*. Springer, Cham, 2021.

[169] Yannik Gerlach, Alexander Seeliger, Timo Nolle, and Max M¨uhlh¨auser. Inferring a multi-perspective likelihood graph from black-box next event predictors. In Xavier Franch, Geert Poels, Frederik Gailly, and Monique Snoeck, editors, *Advanced Information Systems Engineering - 34th International Conference, CAiSE 2022, Leuven, Belgium, June 6-10, 2022, Proceedings*, volume 13295 of *Lecture Notes in Computer Science*, pages 19–35. Springer, 2022.

[170] Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3681–3688. AAAI Press, 2019.

[171] Bill Gillham. *Developing a questionnaire*. A&C Black, 2008.

[172] David M Gligor, Carol L Esmark, and Ismail Gölgeci. Building international business theory: A grounded theory approach. *Journal of International Business Studies*, 47(1):93–111, January 2016.

[173] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual

conditional expectation. *Journal of Computational and Graphical Statistics*, 24, 09 2015.

[174] Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3):50–57, October 2017. Number: 3.

[175] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, John O. Prior, Lode Lauwaert, Wessel Reijers, Adrien Depeursinge, Vincent Andrearczyk, and Henning Müller. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artif. Intell. Rev.*, 56(4):3473–3504, 2023.

[176] Thomas Grisold, Jan Mendling, Markus Otto, and Brocke Jan vom. Adoption, use and management of process mining in practice. *Business Process Management*, 27(2):369–387, 2020.

[177] Michael Grohs, Luka Abb, Nourhan Elsayed, and Jana-Rebecca Rehse. Large language models can accomplish business process management tasks. *CoRR*, abs/2307.09923, 2023.

[178] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[179] Mehmet A. Gulum, Christopher M. Trombley, and Mehmed M. Kantardzic. Multiple interpretations improve deep learning transparency for prostate lesion detection. In Vijay Gadepally, Timothy G. Mattson, Michael Stonebraker, Tim Kraska, Fusheng Wang, Gang Luo, Jun Kong, and Alevtina Dubovitskaya, editors, *Heterogeneous Data Management, Polystores, and Analytics for Healthcare - VLDB Workshops, Poly 2020 and DMAH 2020, Virtual Event, August 31 and September 4, 2020, Revised Selected Papers*, volume 12633 of *Lecture Notes in Computer Science*, pages 120–137. Springer, 2020.

[180] David Gunning and David W. Aha. Darpa's explainable artificial intelligence (XAI) program. *AI Mag.*, 40(2):44–58, 2019.

[181] Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors. *Advances in Neural*

*Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

[182] Khan Mohammad Habibullah and Jennifer Horkoff. Non-functional requirements for machine learning: Understanding current use and challenges in industry. In *29th IEEE International Requirements Engineering Conference, RE 2021, Notre Dame, IN, USA, September 20-24, 2021*, pages 13–23, USA, 2021. IEEE.

[183] Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. Explainable AI under contract and tort law: legal incentives and technical challenges. *Artif Intell Law*, 28(4):415–439, December 2020.

[184] Philipp Hacker and Jan-Hendrik Passoth. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Lecture Notes in Computer Science, pages 343–373, Cham, 2022. Springer International Publishing.

[185] Ulrike Hahn. The problem of circularity in evidence, argument, and explanation. *Perspectives on Psychological Science*, 6(2):172–182, 2011.

[186] Yacov Y Haimes. *Risk modeling, assessment, and management*. John Wiley & Sons, 2015.

[187] Xue Han, Lianxue Hu, Jaydeep Sen, Yabin Dang, Buyu Gao, Vatche Isahagian, Chuan Lei, Vasilis Efthymiou, Fatma Özcan, Abdul Quamar, Ziming Huang, and Vinod Muthusamy. Bootstrapping natural language querying on process automation data. In *2020 IEEE International Conference on Services Computing (SCC)*, pages 170–177, 2020. ISSN: 2474-2473.

[188] Leif Hancox-Li. Robustness in machine learning explanations: does it matter? In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 640–647. ACM, 2020.

[189] Khadijah M. Hanga, Yevgeniya Kovalchuk, and Mohamed Medhat Gaber. A graph-based approach to interpreting recurrent neural networks in process mining. *IEEE Access*, 8:172923–172938, 2020.

[190] M. Harl, S. Weinzierl, M. Stierle, and M. Matzner. Explainable predictive business process monitoring using gated graph neural networks. *Journal of Decision Systems*, 29(sup1):312–327, 2020.

[191] Maximilian Harl, Sven Weinzierl, Matthias Stierle, and Martin Matzner. Explainable predictive business process monitoring using gated graph neural networks. *J. Decis. Syst.*, 29(Supplement):312–327, 2020.

[192] Gilbert H Harman. The inference to the best explanation. *The philosophical review*, 74(1):88–95, 1965.

[193] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009.

[194] Jeffrey Heer. The partnership on ai. *AI Matters*, 4(3):25–26, oct 2018.

[195] Harry Heft. Affordances, dynamic experience, and the challenge of reification. *Ecological psychology*, 15(2):149–180, 2003.

[196] Alice Hein, Lukas J. Meier, Alena Buyx, and Klaus Diepold. A fuzzy-cognitive-maps approach to decision-making in medical ethics. In *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padua, Italy, July 18-23, 2022*, pages 1–8, Padua, Italy, July 18-23, 2022, 2022. IEEE.

[197] B. Heinrichs and S.B. Eickhoff. Your evidence? machine learning algorithms for medical diagnosis and prediction. *Hum Brain Mapp*, 41(6):1435–1444, Apr 15 2020.

[198] Carl G. Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175, 1948.

[199] Christian Herzog. On the ethical and epistemological utility of explicable ai in medicine. *Philosophy and Technology*, 35(2):1–31, 2022.

[200] Frederic Heymans, Thomas Gils, and Wannes Ooms. From policy to practice: Prototyping the eu ai act's transparency requirements. *Available at SSRN 4714345*, February 2024.

[201] Thomas E. Hill. *Dignity and practical reason in Kant's moral theory*. Cornell University Press, 1992.

[202] Lennart Hofeditz, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz. Applying XAI to an ai-based system for candidate management to mitigate bias and discrimination in hiring. *Electron. Mark.*, 32(4):2207–2233, 2022.

[203] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9734–9745, 2019.

[204] Henry A Hornstein. The integration of project management and organizational change management is now a necessity. *International journal of project management*, 33(2):291–298, 2015.

[205] William S Horton and Boaz Keysar. When do speakers take into account common ground? *Cognition*, 59(1):91–117, 1996.

[206] Chihcheng Hsieh, Catarina Moreira, and Chun Ouyang. Dice4el: Interpreting process predictions using a milestone-aware counterfactual approach. In Ciccio et al. [81], pages 88–95.

[207] Chan Hsu, Wei-Chun Huang, Jun-Ting Wu, Chih-Yuan Li, and Yihuang Kang. Toward transparent sequence models with model-based tree markov model. In *6th IEEE International Conference on Multimedia Information Processing and Retrieval, MIPR 2023, Singapore, August 30 - Sept. 1, 2023*, pages 1–6. IEEE, 2023.

[208] Tsung-Hao Huang, Andreas Metzger, and Klaus Pohl. Counterfactual explanations for predictive business process monitoring. In Marinos Themistocleous and Maria Papadaki, editors, *Information Systems - 18th European, Mediterranean, and Middle Eastern Conference, EMCIS 2021, Virtual Event, December 8-9, 2021, Proceedings*, volume 437 of *Lecture Notes in Business Information Processing*, pages 399–413. Springer, 2021.

[209] Tobias Huber, Benedikt Limmer, and Elisabeth André. Benchmarking perturbation-based saliency maps for explaining atari agents. *Frontiers Artif. Intell.*, 5, 2022.

[210] Rosalind Hursthouse. *On virtue ethics*. Oxford University Press, 1999.

[211] Steven E Hyman. The diagnosis of mental disorders: the problem of reification. *Annual review of clinical psychology*, 6:155–179, 2010.

[212] Javier Camacho Ibáñez and Mónica Villas Olmeda. Operationalising ai ethics: how are companies bridging the gap between practice and principles? an exploratory study. *AI & SOCIETY*, 37(4):1663–1687, 2022.

[213] Intelligent Transporation Systems IEEE Vehicular Technology Society (IEEE VT/ITS). Ieee 7001-2021 - standard for transparency of autonomous systems. `https://standards.ieee.org/ieee/7001/6929/`, 2022.

[214] Yoel Inbar and Joris Lammers. Political diversity in social and personality psychology. *Perspectives on Psychological Science*, 7(5):496–503, 2012.

[215] Information Commissioner's Office (ICO) of the United Kingdom and The Alan Turing Institute. Project explain - interim report, 2019.

[216] Information Commissioner's Office (ICO) of the United Kingdom and The Alan Turing Institute. Explaining decisions made with ai, 2020.

[217] JTC 1 /SC 42 Artificial Intelligence International Standards Association (ISO). Iso/iec jtc 1/sc 42 - artificial intelligence committee. `https://www.iso.org/committee/6794475.html`, 2017.

[218] Securing Artificial Intelligence (SAI) International Standards Association (ISO). Iso/iec awi ts 6254 -information technology — artificial intelligence — objectives and approaches for explainability of ml models and ai systems. `https://www.iso.org/standard/82148.html`, 2023.

[219] inVerbis. Healthcare. process mining in a public hospital. *Inverbis*, Jun 2022.

[220] inVerbis. Healthcare management: Performance booster by inverbis. *Inverbis*, Mar 2024.

[221] International Standards Association (ISO). Iso/iec tr 24028:2020 - information technology — artificial intelligence — overview of trustworthiness in artificial intelligence. `https://www.iso.org/standard/77608.html`, 2020.

[222] International Standards Association (ISO). Iso/iec awi 12792 - information technology — artificial intelligence — transparency taxonomy of ai systems. https://www.iso.org/standard/84111.html, 2025.

[223] Jean-Marie John-Mathews. Some critical and ethical perspectives on the empirical turn of AI interpretability. *CoRR*, abs/2109.09586, 2021.

[224] Ian Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374:20150202, 04 2016.

[225] Fleur Jongepier and Esther Keymolen. Explanation and agency: exploring the normative-epistemic landscape of the "right to explanation". *Ethics Inf. Technol.*, 24(4):49, 2022.

[226] Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.

[227] Daniel Kahneman and Amos Tversky. Choices, values, and frames. *American psychologist*, 39(4):341, 1984.

[228] Anna A. Kalenkova, Wil M. P. van der Aalst, Irina A. Lomazova, and Vladimir A. Rubin. Process mining using BPMN: Relating event logs and process models. *Softw Syst Model*, 16(4):1019–1048, 2017-10-01.

[229] Margot E. Kaminski. "The Right to Explanation, Explained". *Berkeley Technology Law Journal*, 2019.

[230] Immanuel Kant. *Foundations of the metaphysics of morals*. Bobbs-Merrill, 1959.

[231] Immanuel Kant. *The metaphysics of morals*. Cambridge University Press, 1996.

[232] Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 228–236. ACM, 2021.

[233] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of CHI 2020*, CHI '20, page 1–14, New York, NY, USA, 2020. ACM.

[234] Mohsen Kazemian and Markus Helfert. A lightweight encryption method for privacy-preserving in process mining. *CoRR*, abs/2304.03579, 2023.

[235] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4466–4474. ijcai.org, 2021.

[236] Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In Ian Watson and Rosina O. Weber, editors, *Case-Based Reasoning Research and Development - 28th International Conference, ICCBR 2020, Salamanca, Spain, June 8-12, 2020, Proceedings*, volume 12311 of *Lecture Notes in Computer Science*, pages 163–178. Springer, 2020.

[237] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

[238] Frank C Keil, Robert Andrew Wilson, and Robert Anton Wilson. *Explanation and cognition*. MIT press, 2000.

[239] Hendrik Kempt, Jan-Christoph Heilinger, and Saskia K. Nagel. Relative explainability and double standards in medical decision-making. *Ethics Inf. Technol.*, 24(2):20, 2022.

[240] Eoin M. Kenny and Mark T. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11575–11585. AAAI Press, 2021.

[241] Boaz Keysar and Bridget Bly. Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language*, 34(1):89–109, 1995.

[242] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. Examples are not enough, learn to criticize! criticism for interpretability. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2280–2288, 2016.

[243] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Dy and Krause [114], pages 2673–2682.

[244] Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Sanity simulations for saliency methods. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 11173–11200. PMLR, 2022.

[245] Tae Wan Kim and Bryan R Routledge. Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Business Ethics Quarterly*, 32(1):75–102, 2022.

[246] Christopher Klinkmüller, Richard Müller, and Ingo Weber. Mining process mining practices: An exploratory characterization of information needs in process analytics. In Thomas Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling, editors, *BPM 2019*, pages 322–337, Cham, 2019. Springer.

[247] Christopher Klinkmüller, Richard Müller, and Ingo Weber. Mining process mining practices: An exploratory characterization of information needs in process analytics. In Thomas T. Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling, editors, *Business Process Management - 17th International Conference, BPM 2019, Vienna, Austria, September 1-6,*

*2019, Proceedings*, volume 11675 of *Lecture Notes in Computer Science*, pages 322–337. Springer, 2019.

[248] Jonghyeon Ko and Marco Comuzzi. A systematic review of anomaly detection for business process event logs. *Bus. Inf. Syst. Eng.*, 65(4):441–462, 2023.

[249] Meriana Kobeissi, Nour Assy, Walid Gaaloul, Bruno Defude, and Bassem Haidar. An intent-based natural language interface for querying process execution data. In Ciccio et al. [81], pages 152–159.

[250] Ron Kohavi. The power of decision tables. In Nada Lavrac and Stefan Wrobel, editors, *Machine Learning: ECML-95, 8th European Conference on Machine Learning, Heraclion, Crete, Greece, April 25-27, 1995, Proceedings*, volume 912 of *Lecture Notes in Computer Science*, pages 174–189. Springer, 1995.

[251] Maximilian A. Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. Explainability as a non-functional requirement. In Daniela E. Damian, Anna Perini, and Seok-Won Lee, editors, *27th IEEE International Requirements Engineering Conference, RE 2019, Jeju Island, Korea (South), September 23-27, 2019*, pages 363–368, Jeju Island, Korea (South), 2019. IEEE.

[252] Jelmer J. Koorn, Iris Beerepoot, Vinicius Stein Dani, Xixi Lu, Inge van de Weerd, Henrik Leopold, and Hajo A. Reijers. Bringing rigor to the qualitative evaluation of process mining findings: An analysis and a proposal. In *2021 3rd International Conference on Process Mining (ICPM)*, pages 120–127, 2021-10.

[253] Christine M. Korsgaard. *Creating the kingdom of ends*. Cambridge University Press, 1996.

[254] Agnes Koschmider, Kay Kaczmarek, Mathias Krause, and Sebastiaan J. van Zelst. Demystifying noise and outliers in event logs: Review and future directions. In Andrea Marrella and Barbara Weber, editors, *Business Process Management Workshops - BPM 2021 International Workshops, Rome, Italy, September 6-10, 2021, Revised Selected Papers*, volume 436 of *Lecture Notes in Business Information Processing*, pages 123–135. Springer, 2021.

[255] John Kotter. Leading change: Why transformation efforts fail. *Harvard business review.*, 85(1), 2007-01-01.

[256] J. Krijger, T. Thuis, M. de Ruiter, E. Ligthart, and I. Broekman. The AI ethics maturity model: a holistic approach to advancing ethical data science in organizations. *AI Ethics*, October 2022.

[257] Jon A Krosnick. Questionnaire design. *The Palgrave handbook of survey research*, pages 439–455, 2018.

[258] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[259] Arie W Kruglanski. The psychology of being" right": The problem of accuracy in social perception and cognition. *Psychological bulletin*, 106(3):395, 1989.

[260] Arie W Kruglanski, Amiram Raviv, Daniel Bar-Tal, Alona Raviv, Keren Sharvit, Shmuel Ellis, and L Mannetti. Says who? epistemic authority effects in social judgment. *Advances in experimental social psychology*, 37:345–392, 2005.

[261] Ulrike Kuhl, André Artelt, and Barbara Hammer. Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2125–2137. ACM, 2022.

[262] Thomas S Kuhn. Objectivity, value judgment, and theory choice. *Arguing about science*, pages 74–86, 1977.

[263] Aditya Kuppa and Nhien-An Le-Khac. Black box attacks on explainable artificial intelligence (xai) methods in cyber security. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[264] Gabriel Laberge, Ulrich Aïvodji, and Satoshi Hara. Fooling SHAP with stealthily biased sampling. *CoRR*, abs/2205.15419, 2022.

[265] Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5628–5638. PMLR, 2020.

[266] Himabindu Lakkaraju and Osbert Bastani. "how do I fool you?": Manipulating user trust via misleading black box explanations. In Markham et al. [306], pages 79–85.

[267] George Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press, 2008.

[268] George Lakoff, Mark Johnson, and John F Sowa. Review of philosophy in the flesh: The embodied mind and its challenge to western thought. *Computational Linguistics*, 25(4):631–634, 1999.

[269] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, July 2021.

[270] Stefan Larsson and Fredrik Heintz. Transparency in artificial intelligence. *Internet Policy Rev.*, 9(2), 2020.

[271] Katsiaryna Lashkevich, Lino Moises Mediavilla Ponce, Manuel Camargo, Fredrik Milani, and Marlon Dumas. Discovery of improvement opportunities in knock-out checks of business processes. In Selmin Nurcan, Andreas L. Opdahl, Haralambos Mouratidis, and Aggeliki Tsohou, editors, *Research Challenges in Information Science: Information Science and the Connected World - 17th International Conference, RCIS 2023, Corfu, Greece, May 23-26, 2023, Proceedings*, volume 476 of *Lecture Notes in Business Information Processing*, pages 381–397. Springer, 2023.

[272] Sander J. J. Leemans, Dirk Fahland, and Wil M. P. van der Aalst. Discovering block-structured process models from event logs - A constructive approach. In José Manuel Colom and Jörg Desel, editors, *Application and Theory of Petri Nets and Concurrency - 34th International Conference, PETRI NETS 2013, Milan, Italy, June 24-28, 2013. Proceedings*, volume 7927 of *Lecture Notes in Computer Science*, pages 311–329. Springer, 2013.

[273] Sander J. J. Leemans, Dirk Fahland, and Wil M. P. van der Aalst. Discovering block-structured process models from event logs containing infrequent behaviour. In Niels Lohmann, Minseok Song, and Petia Wohed, editors, *Business Process Management Workshops - BPM 2013 International Workshops,*

*Beijing, China, August 26, 2013, Revised Papers*, volume 171 of *Lecture Notes in Business Information Processing*, pages 66–78. Springer, 2013.

[274] Teemu Lehto and Markku Hinkka. Discovering business area effects to process mining analysis using clustering and influence analysis. In Witold Abramowicz and Gary Klein, editors, *Business Information Systems - 23rd International Conference, BIS 2020, Colorado Springs, CO, USA, June 8-10, 2020, Proceedings*, volume 389 of *Lecture Notes in Business Information Processing*, pages 236–248. Springer, 2020.

[275] Giorgio Leonardi, Stefania Montani, and Manuel Striani. Explainable process trace classification: An application to stroke. *J. Biomed. Informatics*, 126:103981, 2022.

[276] Henrik Leopold, Jan Mendling, and Artem Polyvyanyy. Supporting process model validation through natural language generation. *IEEE Transactions on Software Engineering*, 40(8):818–840, 2014.

[277] Anastasia-M Leventi-Peetz and Kai Weber. Rashomon effect and consistency in explainable artificial intelligence (xai). In *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1*, pages 796–808. Springer, 2022.

[278] Sergej Levich, Bernhard Lutz, and Dirk Neumann. Utilizing the omnipresent: Incorporating digital documents into predictive process monitoring using deep neural networks. *Decis. Support Syst.*, 175:114043, 2023.

[279] Stephen C Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

[280] David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In Claire Nedellec and Céline Rouveirol, editors, *Machine Learning: ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany, April 21-23, 1998, Proceedings*, volume 1398 of *Lecture Notes in Computer Science*, pages 4–15. Springer, 1998.

[281] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In McIlraith and Weinberger [312], pages 3530–3537.

[282] Gabriel Lima, Nina Grgic-Hlaca, Jin Keun Jeong, and Meeyoung Cha. The conflict between explainable and accountable decision-making algorithms. In

*FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 2103–2113, Jeju Island, Korea (South), 2022. ACM.

[283] Felix Lindner and Katrin Möllney. Extracting reasons for moral judgments under various ethical principles. In Christoph Benzmüller and Heiner Stuckenschmidt, editors, *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, volume 11793 of *Lecture Notes in Computer Science*, pages 216–229, Germany, 2019. Springer.

[284] Peter Lipton. Inference to the best explanation. *A Companion to the Philosophy of Science*, pages 184–193, 2017.

[285] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, jun 2018.

[286] Ao Liu, Xiaoyu Chen, Sijia Liu, Lirong Xia, and Chuang Gan. Certifiably robust interpretation via rényi differential privacy. *Artif. Intell.*, 313:103787, 2022.

[287] De Liu, Radhika Santhanam, and Jane Webster. Toward meaningful engagement: A framework for design and research of gamified information systems. *MIS Q.*, 41(4):1011–1034, 2017.

[288] Edwin A Locke and Gary P Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705, 2002.

[289] Edwin A Locke and Gary P Latham. New directions in goal-setting theory. *Current directions in psychological science*, 15(5):265–268, 2006.

[290] Helena Löfström, Karl Hammar, and Ulf Johansson. A meta survey of quality evaluation criteria in explanation methods. In Weerdt and Polyvyanyy [494], pages 55–63.

[291] Tania Lombrozo. The instrumental value of explanations. *Philosophy Compass*, 6(8):539–551, 2011.

[292] Tania Lombrozo. Explanation and abductive inference. *The Oxford Handbook of Thinking and Reasoning*, 2012.

[293] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.

[294] Iezalde F. Lopes and Diogo R. Ferreira. A survey of process mining competitions: The BPI challenges 2011–2018. In Chiara Di Francescomarino, Remco Dijkman, and Uwe Zdun, editors, *Business Process Management Workshops*, volume 362, pages 263–274. Springer International Publishing, 2019. Series Title: Lecture Notes in Business Information Processing.

[295] Scott M. Lundberg, Gabriel G. Erion, Hugh Chen, Alex J. DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, 2(1):56–67, 2020.

[296] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Guyon et al. [181], pages 4765–4774.

[297] P. M. Dixit, H. S. Garcia Caballero, A. Corvò, B. F. A. Hompes, J. C. A. M. Buijs, and W. M. P. van der Aalst. Enabling interactive process analysis with process mining and visual analytics:. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 573–584. SCITEPRESS - Science and Technology Publications, 2017.

[298] Alasdair MacIntyre. *After virtue: A study in moral theory*. University of Notre Dame Press, 1981.

[299] Fabrizio Maria Maggi, Chiara Di Francescomarino, Marlon Dumas, and Chiara Ghidini. Predictive monitoring of business processes. In Matthias Jarke, John Mylopoulos, Christoph Quix, Colette Rolland, Yannis Manolopoulos, Haralambos Mouratidis, and Jennifer Horkoff, editors, *Advanced Information Systems Engineering - 26th International Conference, CAiSE 2014, Thessaloniki,*

*Greece, June 16-20, 2014. Proceedings*, volume 8484 of *Lecture Notes in Computer Science*, pages 457–472. Springer, 2014.

[300] Ana Rocío Cárdenas Maita, Marcelo Fantinato, Sarajane Marques Peres, and Fabrizio Maria Maggi. Towards a business-oriented approach to visualization-supported interpretability of prediction results in process mining. In Joaquim Filipe, Michal Smialek, Alexander Brodsky, and Slimane Hammoudi, editors, *Proceedings of the 25th International Conference on Enterprise Information Systems, ICEIS 2023, Volume 1, Prague, Czech Republic, April 24-26, 2023*, pages 395–406. SCITEPRESS, 2023.

[301] Gianclaudio Malgieri. Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Computer Law & Security Review*, 35(5):105327, October 2019.

[302] Gianclaudio Malgieri and Frank Pasquale. Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. *Computer Law & Security Review*, 52:105899, 2024.

[303] Felix Mannhardt. Multi-perspective process mining. In Wil M. P. van der Aalst, Fabio Casati, Raffaele Conforti, Massimiliano de Leoni, Marlon Dumas, Akhil Kumar, Jan Mendling, Surya Nepal, Brian T. Pentland, and Barbara Weber, editors, *Proceedings of the Dissertation Award, Demonstration, and Industrial Track at BPM 2018 co-located with 16th International Conference on Business Process Management (BPM 2018), Sydney, Australia, September 9-14, 2018*, volume 2196 of *CEUR Workshop Proceedings*, pages 41–45. CEUR-WS.org, 2018.

[304] Felix Mannhardt. Responsible process mining. In van der Aalst and Carmona [469], pages 373–401.

[305] Ronny Mans, Wil M. P. van der Aalst, and Rob J. B. Vanwersch. *Process Mining in Healthcare - Evaluating and Exploiting Operational Healthcare Processes*. Springer Briefs in Business Process Management. Springer, 2015.

[306] Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington, editors. *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*. ACM, 2020.

[307] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J. Biomed. Informatics*, 113, 2021.

[308] Niels Martin, Dominik A Fischer, Georgi D Kerpedzhiev, Kanika Goel, Sander JJ Leemans, Maximilian Röglinger, Wil MP van der Aalst, Marlon Dumas, Marcello La Rosa, and Moe T Wynn. Opportunities and challenges for process mining in organizations: results of a delphi study. *Bus Inf Syst Eng*, 63:511–527, 2021.

[309] Andreia Martinho, Maarten Kroesen, and Caspar G. Chorus. A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence. *Artif. Intell. Medicine*, 121:102190, 2021.

[310] Yoshihiro Maruyama. Categorical artificial intelligence: The integration of symbolic and statistical AI for verifiable, ethical, and trustworthy AI. In Ben Goertzel, Matthew Iklé, and Alexey Potapov, editors, *Artificial General Intelligence - 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15-18, 2021, Proceedings*, volume 13154 of *Lecture Notes in Computer Science*, pages 127–138. Springer, 2021.

[311] Craig Ed McGarty, Vincent Y Yzerbyt, and Russell Ed Spears. *Stereotypes as Explanations: The Formation of Meaningful Beliefs about Social Groups*. Cambridge University Press, 2002.

[312] Sheila A. McIlraith and Kilian Q. Weinberger, editors. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press, 2018.

[313] Nijat Mehdiyev and Peter Fettke. *Explainable Artificial Intelligence for Process Mining: A General Overview and Application of a Novel Local Explanation Approach for Predictive Process Monitoring*, pages 1–28. Springer International Publishing, Cham, 2021.

[314] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8930–8938, May 2021.

[315] Jörg Meibauer. Tautology as presumptive meaning. *Pragmatics & cognition*, 16(3):439–470, 2008.

[316] Jan Mendling, Hajo A. Reijers, and Jorge Cardoso. What makes process models understandable? In Gustavo Alonso, Peter Dadam, and Michael Rosemann, editors, *Business Process Management*, volume 4714, pages 48–63. Springer Berlin Heidelberg, 2007. Series Title: Lecture Notes in Computer Science.

[317] Jan Mendling, Hajo A. Reijers, and William M.P. van der Aalst. Seven process modeling guidelines (7PMG). *Information and Software Technology*, 52(2):127–136, 2010-02.

[318] Jan Mendling, Mark Strembeck, and Jan Recker. Factors of process model comprehension—findings from a series of experiments. *Decision Support Systems*, 53(1):195–206, 2012.

[319] Rosa Meo, Roberto Nai, and Emilio Sulis. Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next? In Silvia Chiusano, Tania Cerquitelli, and Robert Wrembel, editors, *Advances in Databases and Information Systems - 26th European Conference, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings*, volume 13389 of *Lecture Notes in Computer Science*, pages 25–34, Italy, 2022. Springer.

[320] Erwan Le Merrer and Gilles Trédan. Remote explainability faces the bouncer problem. *Nat. Mach. Intell.*, 2(9):529–539, 2020.

[321] Jacob Metcalf and Kate Crawford. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211, 2016.

[322] Ryszard S. Michalski. A theory and methodology of inductive learning. *Artif. Intell.*, 20(2):111–161, 1983.

[323] Alexandra Michel, Rune Todnem By, and Bernard Burnes. The limitations of dispositional resistance in relation to organizational change. *Management Decision*, 51:761–780, 2013.

[324] John Stuart Mill. *Utilitarianism*. Hackett Publishing, 1979.

[325] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[326] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

[327] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *CoRR*, abs/2111.00358, 2021.

[328] Ronald K Mitchell, Bradley R Agle, and Donna J Wood. Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of management review*, 22(4):853–886, 1997.

[329] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst.*, 11(3-4):24:1–24:45, 2021.

[330] Jakob Mökander and Luciano Floridi. Operationalising ai governance through ethics-based auditing: an industry case study. *AI and Ethics*, pages 1–18, 2022.

[331] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15, 2018.

[332] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. Operationalising AI ethics: barriers, enablers and next steps. *AI Soc.*, 38(1):411–423, 2023.

[333] Miranda X Morris, Ethan Y Song, Aashish Rajesh, Malke Asaad, and Brett T Phillips. Ethical, legal, and financial considerations of artificial intelligence in surgery. *The American Surgeon*, 89(1):55–60, 2023.

[334] Azadeh Sadat Mozafari Mehr, Renata M. de Carvalho, and Boudewijn van Dongen. Explainable conformance checking: Understanding patterns of anomalous behavior. *Eng. Appl. Artif. Intell.*, 126(PB), feb 2024.

[335] Markus Mueck, Raymond Forbes, Scott Cadzow, Suno Wood, and European Telecommunications Standards Institute (ETSI) Gazis, Evangelos. Etsi activities in the field of artificial intelligence - preparing the implementation of the european ai act, 2022.

[336] Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):119:1–119:36, 2019.

[337] Jorge Munoz-Gama and Xixi Lu, editors. *Process Mining Workshops - ICPM 2021 International Workshops, Eindhoven, The Netherlands, October 31 - November 4, 2021, Revised Selected Papers*, volume 433 of *Lecture Notes in Business Information Processing*. Springer, 2022.

[338] Ulla-Maija Mylly. Transparent ai? navigating between rules on trade secrets and access to information. *IIC - International Review of Intellectual Property and Competition Law*, 54(7):1013–1043, 2023.

[339] Karim Nadim, Mohamed-Salah Ouali, Hakim Ghezzaz, and Ahmed Ragab. Learn-to-supervise: Causal reinforcement learning for high-level control in industrial processes. *Engineering Applications of Artificial Intelligence*, 126:106853, 2023.

[340] Karim Nadim, Ahmed Ragab, and Mohamed-Salah Ouali. Data-driven dynamic causality analysis of industrial systems using interpretable machine learning and process mining. *J. Intell. Manuf.*, 34(1):57–83, 2023.

[341] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 466–477, New York, NY, USA, 2021. Association for Computing Machinery.

[342] Luca Nannini, Jose Maria Alonso-Moral, Alejandro Catala, Manuel Lama, and Senén Barro. Operationalizing Explainable AI in the EU Regulatory Ecosystem. *IEEE Intelligent Systems*, Forthcoming(01):1–13, apr 2024.

[343] Luca Nannini, Agathe Balayn, and Adam Leon Smith. Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the eu, us, and UK. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 1198–1212. ACM, 2023.

[344] Devesh Narayanan and Zhi Ming Tan. Attitudinal tensions in the joint pursuit of explainable and trusted AI. *Minds Mach.*, 33(1):55–82, 2023.

[345] Keith Ng and Stewart Hase. Grounded suggestions for doing a grounded theory business research. *Electronic Journal of Business Research Methods*, 6(2):pp183–198, 2008.

[346] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[347] Claire Nicodeme. Build confidence and acceptance of ai-based decision support systems - explainable and liable AI. In *13th International Conference on Human System Interaction, HSI 2020, Tokyo, Japan, June 6-8, 2020*, pages 20–23. IEEE, 2020.

[348] Adam Noack, Isaac Ahern, Dejing Dou, and Boyang Li. An empirical study on the relation between network interpretability and adversarial robustness. *SN Comput. Sci.*, 2(1):32, 2021.

[349] Maximilian Noppel, Lukas Peter, and Christian Wressnegger. Disguising attacks with explanation-aware backdoors. In *2023 2023 IEEE Symposium on Security and Privacy (SP) (SP)*, pages 664–681, Los Alamitos, CA, USA, may 2023. IEEE Computer Society.

[350] Claudia Novelli, Philipp Hacker, Jessica Morley, Jarle Trondal, and Luciano Floridi. A robust governance for the ai act: Ai office, ai board, scientific panel, and national authorities. *Available at SSRN*, 2024.

[351] Martha Nussbaum. Non-relative virtues: an aristotelian approach. *Midwest studies in philosophy*, 13(1):32–53, 1988.

[352] Justin Oakley. Varieties of virtue ethics. *Ratio*, 9(2):128–152, 1996.

[353] IMCO-LIBE Committees of European Parliament. Draft report on the proposal for a regulation of the european parliament and of the council on harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2022-04-20.

[354] JURI Committee of European Parliament. Opinion of the committee on legal affairs for the committee on the internal market and consumer protection and the committee on civil liberties, justice and home affairs on the proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2022-09-12.

[355] Office of Science and Technology Policy of the White House of United States of America. Blueprint for an ai bill of rights: Making automated systems work for the american people, 2022.

[356] National Institute of Standards and Technology (NIST) of United States of America. U.s. leadership in ai: A plan for federal engagement in developing technical standards and related tools - prepared in response to executive order 13859 submitted on august 9, 2019, 2021.

[357] National Institute of Standards and Technology (NIST) of United States of America. Risk management framework v1.0, 2023.

[358] Commission of the European Union. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021-04-21.

[359] Commission of the European Union. Proposal for a directive of the european parliament and of the council on adapting non-contractual civil liability rules to artificial intelligence (ai liability directive), 2022-09-28.

[360] Council of the European Union. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts - general approach, 2022-11-25.

[361] Executive Office of the President of United States of America. Maintaining american leadership in artificial intelligence - executive order 13859 of february 11, 2019, 2019.

[362] Executive Office of the President of United States of America Office of Management and Budget. Guidance for regulation of artificial intelligence applications, 2020.

[363] Information Commissioner's Office (ICO) of the United Kingdom. Guidance on ai and data protection, 2020.

[364] Information Commissioner's Office (ICO) of the United Kingdom. Guidance on the ai auditing framework: Draft guidance for consultation, 2020.

[365] Parliament of the United Kingdom. Data protection act 2018, 2018.

[366] Defense Advanced Research Projects Agency of United States of America. Federal contract opportunity for explainable artificial intelligence (xai) darpa-baa-16-53, 2016-08-050.

[367] Cabinet Office & Central Digital & Data Office & Office for Artificial Intelligence. Ethics, transparency and accountability framework for automated decision-making - guidance, 2021.

[368] Seong Joon Oh, Bernt Schiele, and Mario Fritz. Towards reverse-engineering black-box neural networks. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144, 2019.

[369] High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI, 2019.

[370] High-Level Expert Group on Artificial Intelligence. The assessment list for trustworthy artificial intelligence (altai), 2020.

[371] National Security Commission on Artificial Intelligence of United States of America. Final report, 2021.

[372] Onora O'Neill. *Acting on principle: An essay on Kantian ethics*. Columbia University Press, 1975.

[373] Charles A. O'Reilly. Individuals and information overload in organizations: Is more necessarily better? *Academy of Management Journal*, 23(4):684–696, 1980.

[374] Alessandro Padella, Massimiliano de Leoni, Onur Dogan, and Riccardo Galanti. Explainable process prescriptive analytics. In Burattin et al. [67], pages 16–23.

[375] Tiago Palma Pagano, Rafael Bessa Loureiro, Fernanda Vitoria Nascimento Lisboa, Rodrigo Matos Peixoto, Guilherme Aragao de Sousa Guimaraes, Gustavo Oliveira Ramos Cruz, Maira Matos Araujo, Lucas Lisboa dos Santos, Marco Cruz, Ewerton Lopes Silva de Oliveira, Ingrid Winkler, and Erick Giovani Sperandio Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data Cogn. Comput.*, 7(1):15, 2023.

[376] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

[377] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernández Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sánchez, Josep Soler Garrido, and Emilia Gómez. The role of explainable AI in the context of the AI act. In *FAccT*, pages 1139–1150. ACM, 2023.

[378] Derek Parfit. *Reasons and persons*. Oxford University Press, 1984.

[379] Gyunam Park, Aaron K¨usters, Mara Tews, Cameron Pitsch, Jonathan Schneider, and Wil M. P. van der Aalst. Explainable predictive decision mining for operational support. In Javier Troya, Raffaela Mirandola, Elena Navarro, Andrea Delgado, Sergio Segura, Guadalupe Ortiz, Cesare Pautasso, Christian Zirpins, Pablo Fernández, and Antonio Ruiz-Cortés, editors, *Service-Oriented Computing - ICSOC 2022 Workshops - ASOCA, AI-PA, FMCIoT, WESOACS 2022, Sevilla, Spain, November 29 - December 2, 2022 Proceedings*, volume 13821 of *Lecture Notes in Computer Science*, pages 66–79. Springer, 2022.

[380] Parliament and Council of the European Union. Directive (eu) 2016/943 of the european parliament and of the council of 8 june 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure (text with eea relevance). *Official Journal*, 2016-06-15.

[381] Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, and Donato Malerba. Jarvis: Joining adversarial training with vision transformers in next-activity prediction. *IEEE Transactions on Services Computing*, -(01):1–14, nov 2024.

[382] Vincenzo Pasquadibisceglie, Annalisa Appice, Giuseppe Ieva, and Donato Malerba. Tsunami - an explainable ppm approach for customer churn prediction in evolving retail data environments. *J Intell Inf Syst*, 2023.

[383] Vincenzo Pasquadibisceglie, Giovanna Castellano, Annalisa Appice, and Donato Malerba. FOX: a neuro-fuzzy model for process outcome prediction and explanation. In Ciccio et al. [81], pages 112–119.

[384] Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1895–1904, 2022.

[385] Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.

[386] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

[387] P. Jonathon Phillips, Carina Hahn, Peter Fontana, Amy Yates, Kristen K. Greene, David Broniatowski, and Mark A. Przybocki. Four principles of explainable artificial intelligence, 2021-09-29 04:09:00 2021.

[388] Artem Polyvyanyy, Moe Thandar Wynn, Amy Van Looy, and Manfred Reichert, editors. *Business Process Management - 19th International Conference, BPM 2021, Rome, Italy, September 06-10, 2021, Proceedings*, volume 12875 of *Lecture Notes in Computer Science*. Springer, 2021.

[389] Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. routledge, 2014.

[390] Mahsa Pourbafrani, Shreya Kar, Sebastian Kaiser, and Wil M. P. van der Aalst. Remaining time prediction for processes with inter-case dynamics. In Munoz-Gama and Lu [337], pages 140–153.

[391] Pengrui Quan, Supriyo Chakraborty, Jeya Vikranth Jeyakumar, and Mani B. Srivastava. On the amplification of security and privacy risks by post-hoc explanations in machine learning models. *CoRR*, abs/2206.14004, 2022.

[392] J. Ross Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, 1986.

[393] Majid Rafiei and Wil M. P. van der Aalst. Privacy-preserving data publishing in process mining. In Fahland et al. [138], pages 122–138.

[394] Majid Rafiei and Wil M. P. van der Aalst. An abstraction-based approach for privacy-aware federated process mining. *IEEE Access*, 11:33697–33714, 2023.

[395] Majid Rafiei and Wil MP van der Aalst. Towards quantifying privacy in process mining. In *International Conference on Process Mining*, pages 385–397. Springer, 2020.

[396] Peter Railton. Alienation, consequentialism, and the demands of morality. *Philosophy & Public Affairs*, pages 134–171, 1984.

[397] Inioluwa Deborah Raji and Jingying Yang. ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles. *arXiv:1912.06166 [cs, stat]*, January 2020. arXiv: 1912.06166.

[398] Jana-Rebecca Rehse, Nijat Mehdiyev, and Peter Fettke. Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory. *K¨unstliche Intell.*, 33(2):181–187, 2019.

[399] Lars Reinkemeyer. *Process mining in action*. Springer, 2020.

[400] Lars Reinkemeyer. Status and future of process mining: From process discovery to process execution. In van der Aalst and Carmona [469], pages 405–415.

[401] Marjorie Rhodes and Kelsey Moty. What is social essentialism and how does it develop? *Advances in child development and behavior*, 59:1–30, 2020.

[402] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[403] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In McIlraith and Weinberger [312], pages 1527–1535.

[404] Williams Rizzi, Chiara Di Francescomarino, and Fabrizio Maria Maggi. Explainability in predictive process monitoring: When understanding helps improving. In Fahland et al. [138], pages 141–158.

[405] Scott Robbins. A misdirected principle with a catch: Explicability for AI. *Minds Mach.*, 29(4):495–514, 2019.

[406] Raphael De Almeida Rodrigues, Leonardo Guerreiro Azevedo, and Kate Cerqueira Revoredo. BPM2text: A language independent framework for business process models to natural language text. *iSys - Brazilian Journal of Information Systems*, 9(4):38–56, 2016.

[407] Toke Ronnow-Rasmussen. Intrinsic and extrinsic value. In *The Oxford handbook of value theory*, pages 29–43. Oxford University Press, 2015.

[408] William David Ross. *The right and the good*. Clarendon Press, 1930.

[409] Sarah Rossnan. Overcoming math anxiety. *Mathitudes*, 1(1):1–4, 2006.

[410] Martin Rowlson. *Uber: Process Mining to Optimize Customer Experience and Business Performance*, pages 59–63. Springer International Publishing, Cham, 2020.

[411] Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562, 2002.

[412] Anne Rozinat and Wil M. P. van der Aalst. Conformance checking of processes based on monitoring real behavior. *Inf. Syst.*, 33(1):64–95, 2008.

[413] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[414] Waddah Saeed and Christian W. Omlin. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl. Based Syst.*, 263:110273, 2023.

[415] Wesley C Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[416] Wesley C Salmon. Four decades of scientific explanation. *Minnesota studies in the philosophy of science*, 13, 1989.

[417] Josep Sànchez-Ferreres, Andrea Burattin, Josep Carmona, Marco Montali, and Lluís Padró. Formal reasoning on natural language descriptions of processes. In Thomas T. Hildebrandt, Boudewijn F. van Dongen, Maximilian Röglinger, and Jan Mendling, editors, *Business Process Management - 17th International Conference, BPM 2019, Vienna, Austria, September 1-6, 2019, Proceedings*,

volume 11675 of *Lecture Notes in Computer Science*, pages 86–101. Springer, 2019.

[418] Josep Sànchez-Ferreres, Andrea Burattin, Josep Carmona, Marco Montali, Lluís Padró, and Luis Quishpi. Unleashing textual descriptions of business processes. *Softw. Syst. Model.*, 20(6):2131–2153, 2021.

[419] Denise Maria Vecino Sato, Sheila Cristiana De Freitas, Jean Paul Barddal, and Edson Emílio Scalabrin. A survey on concept drift in process mining. *ACM Comput. Surv.*, 54(9):189:1–189:38, 2022.

[420] Roger C Schank. *Making minds less well educated than our own*. Routledge, 2004.

[421] Samuel Scheffler. *The rejection of consequentialism*. Oxford University Press, 7 edition, 1982.

[422] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 617–626, New York, NY, USA, 2022. Association for Computing Machinery.

[423] John R Searle. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, 1979.

[424] Philip Sedgwick. Non-response bias versus response bias. *British Medical Journal*, 348, 2014.

[425] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017.

[426] Amartya Sen. Utilitarianism and welfarism. *The Journal of Philosophy*, 76(9):463–489, 1979.

[427] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K.

Mulligan, editors, *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 231–241. ACM, 2021.

[428] Roee Shraga, Avigdor Gal, Dafna Schumacher, Arik Senderovich, and Matthias Weidlich. Process discovery with context-aware process trees. *Inf. Syst.*, 106:101533, 2022.

[429] Fadi N. Sibai. AI crimes: A classification. In *2020 International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2020, Dublin, Ireland, June 15-19, 2020*, pages 1–8, Ireland, 2020. IEEE.

[430] Henry Sidgwick. *The methods of ethics*. Hackett Publishing, 7 edition, 1907.

[431] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

[432] Renuka Sindhgatta, Chun Ouyang, and Catarina Moreira. Exploring interpretability for predictive process analytics. In Eleanna Kafeza, Boualem Benatallah, Fabio Martinelli, Hakim Hacid, Athman Bouguettaya, and Hamid Motahari, editors, *Service-Oriented Computing - 18th International Conference, ICSOC 2020, Dubai, United Arab Emirates, December 14-17, 2020, Proceedings*, volume 12571 of *Lecture Notes in Computer Science*, pages 439–447. Springer, 2020.

[433] Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing inputs for fragile interpretations in deep natural language processing. In Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021*, pages 420–434. Association for Computational Linguistics, 2021.

[434] Sanchit Sinha, Mengdi Huai, Jianhui Sun, and Aidong Zhang. Understanding and enhancing robustness of concept-based models. *CoRR*, abs/2211.16080, 2022.

[435] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 62–75, 2021.

[436] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In Markham et al. [306], pages 180–186.

[437] Michael Slote. *From morality to virtue*. Oxford University Press, 1992.

[438] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[439] Algorithmic Bias Working Group IEEE Computer Society (IEEE C/S2ESC/ALGB-WG) Software & Systems Engineering Standards Committee. Ieee p7003 - algorithmic bias considerations, 2017.

[440] David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In Frank Hutter, Kristian Kersting, Jefrey Lijffijt, and Isabel Valera, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part I*, volume 12457 of *Lecture Notes in Computer Science*, pages 162–177. Springer, 2020.

[441] Marc Solé and Josep Carmona. Process mining from a basis of state regions. In Johan Lilius and Wojciech Penczek, editors, *Applications and Theory of Petri Nets, 31st International Conference, PETRI NETS 2010, Braga, Portugal, June 21-25, 2010. Proceedings*, volume 6128 of *Lecture Notes in Computer Science*, pages 226–245. Springer, 2010.

[442] Riad Sonbol, Ghaida Rebdawi, and Nada Ghneim. A machine translation like approach to generate business process model from textual description. *SN Comput. Sci.*, 4(3):291, 2023.

[443] Christian Sonnenberg and Jan vom Brocke. Evaluations in the science of the artificial - reconsidering the build-evaluate pattern in design science research.

In Ken Peffers, Marcus A. Rothenberger, and William L. Kuechler Jr., editors, *Design Science Research in Information Systems. Advances in Theory and Practice - 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings*, volume 7286 of *Lecture Notes in Computer Science*, pages 381–397. Springer, 2012.

[444] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. Metrics, Explainability and the European AI Act Proposal. *J*, 5:126–138, February 2022.

[445] Sabin Srivannaboon and Dragan Z. Milosevic. A two-way influence between business strategy and project management. *International Journal of Project Management*, 24(6):493–505, 2006.

[446] P Kyle Stanford. *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*, volume 1. Oxford University Press, 2006.

[447] Ilia Stepin, José Maria Alonso, Alejandro Catalá, and Martin Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[448] Alexander Stevens, Johannes De Smedt, and Jari Peeperkorn. Quantifying explainability in outcome-oriented predictive process monitoring. In Munoz-Gama and Lu [337], pages 194–206.

[449] Alexander Stevens, Johannes De Smedt, Jari Peeperkorn, and Jochen De Weerdt. Assessing the robustness in predictive process monitoring through adversarial attacks. In Burattin et al. [67], pages 56–63.

[450] Matthias Stierle, Jens Brunk, Sven Weinzierl, Sandra Zilker, Martin Matzner, and Jörg Becker. Bringing light into the darkness - A systematic literature review on explainable predictive business process monitoring techniques. In Frantz Rowe, Redouane El Amrani, Moez Limayem, Sabine Matook, Christoph Rosenkranz, Edgar A. Whitley, and Ali El Quammah, editors, *29th European Conference on Information Systems - Human Values Crisis in a Digitizing World, ECIS 2021, Marrakech, Morocco, 2020*, 2021.

[451] Klaas-Jan Stol, Paul Ralph, and Brian Fitzgerald. Grounded theory in software engineering research: A critical review and guidelines. In *Proceedings of*

*the 38th International Conference on Software Engineering*, ICSE '16, page 120–131, New York, NY, USA, 2016. Association for Computing Machinery.

[452] Emily Sullivan and Philippe Verreault-Julien. From explanation to recommendation: Ethical standards for algorithmic recourse. In Vincent Conitzer, John Tasioulas, Matthias Scheutz, Ryan Calo, Martina Mara, and Annette Zimmermann, editors, *AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021*, pages 712–722. ACM, 2022.

[453] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017.

[454] Suriadi Suriadi, Chun Ouyang, Wil M. P. van der Aalst, and Arthur H. M. ter Hofstede. Root cause analysis with enriched process logs. In Marcello La Rosa and Pnina Soffer, editors, *Business Process Management Workshops - BPM 2012 International Workshops, Tallinn, Estonia, September 3, 2012. Revised Papers*, volume 132 of *Lecture Notes in Business Information Processing*, pages 174–186. Springer, 2012.

[455] Josep Sànchez-Ferreres, Han van der Aa, Josep Carmona, and Lluís Padró. Aligning textual and model-based process descriptions. *Data & Knowledge Engineering*, 118:25–40, 2018.

[456] Ruixiang Tang, Ninghao Liu, Fan Yang, Na Zou, and Xia Hu. Defense against explanation manipulation. *Frontiers Big Data*, 5:704203, 2022.

[457] Julian Theis and Houshang Darabi. Decay replay mining to predict next process events. *IEEE Access*, 7:119787–119803, 2019.

[458] Mark Theunissen and Jacob Browning. Putting explainable AI in context: institutional explanations for medical AI. *Ethics Inf. Technol.*, 24(2):23, 2022.

[459] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun D. Preece. Sanity checks for saliency metrics. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI*

*Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6021–6029. AAAI Press, 2020.

[460] Eric Trist. The evolution of socio-technical systems. *Occasional paper*, 2:1981, 1981.

[461] J. D. Trout. Scientific explanation and the sense of understanding. *Philosophy of Science*, 69(2):212–233, 2002.

[462] Richard M. Tubbs, William F. Messier, and W. Robert Knechel. Recency effects in the auditor's belief-revision process. *The Accounting Review*, 65(2):452–460, 1990.

[463] Matteo Turilli and Luciano Floridi. The ethics of information transparency. *Ethics & IT*, 11(2):105–112, 2009.

[464] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

[465] Han van der Aa, Claudio Di Ciccio, Henrik Leopold, and Hajo A. Reijers. Extracting declarative process models from natural language. In Paolo Giorgini and Barbara Weber, editors, *Advanced Information Systems Engineering - 31st International Conference, CAiSE 2019, Rome, Italy, June 3-7, 2019, Proceedings*, volume 11483 of *Lecture Notes in Computer Science*, pages 365–382. Springer, 2019.

[466] Han van der Aa, Henrik Leopold, and Hajo A. Reijers. Checking process compliance against natural language specifications using behavioral spaces. *Information Systems*, 78:83–95, 2018.

[467] Wil van der Aalst. *Process Mining*. Springer Berlin Heidelberg, 2016.

[468] Wil M. P. van der Aalst. Process mining: A 360 degree overview. In van der Aalst and Carmona [469], pages 3–34.

[469] Wil M. P. van der Aalst and Josep Carmona, editors. *Process Mining Handbook*, volume 448 of *Lecture Notes in Business Information Processing*. Springer, 2022.

[470] Wil M. P. van der Aalst, Ton Weijters, and Laura Maruster. Workflow mining: Discovering process models from event logs. *IEEE Trans. Knowl. Data Eng.*, 16(9):1128–1142, 2004.

[471] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[472] Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers Robotics AI*, 8:640647, 2021.

[473] Jan Martijn E. M. van der Werf, Boudewijn F. van Dongen, Cor A. J. Hurkens, and Alexander Serebrenik. Process discovery using integer linear programming. *Fundam. Informaticae*, 94(3-4):387–412, 2009.

[474] Boudewijn van Dongen. Bpi challenge 2019, 2019.

[475] Maikel L. van Eck, Natalia Sidorova, and Wil M. P. van der Aalst. Discovering and exploring state-based models for multi-perspective processes. In Marcello La Rosa, Peter Loos, and Oscar Pastor, editors, *Business Process Management - 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, September 18-22, 2016. Proceedings*, volume 9850 of *Lecture Notes in Computer Science*, pages 142–157. Springer, 2016.

[476] Martijn van Otterlo and Martin Atzmueller. A conceptual view on the design and properties of explainable AI systems for legal settings. In Víctor Rodríguez-Doncel, Monica Palmirani, Michal Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor, editors, *AI Approaches to the Complexity of Legal Systems XI-XII - AICOL International Workshops 2018 and 2020: AICOL-XI@JURIX 2018, AICOL-XII@JURIX 2020, XAILA@JURIX 2020, Revised Selected Papers*, volume 13048 of *Lecture Notes in Computer Science*, pages 143–153, Luxembourg, 2020. Springer.

[477] Sebastiaan J. van Zelst, Felix Mannhardt, Massimiliano de Leoni, and Agnes Koschmider. Event abstraction in process mining: literature review and taxonomy. *Granular Computing*, 6(3):719–736, 2021.

[478] Frederic Vandenberghe. Reification: History of the concept. *International Encyclopedia of the Social and Behavioral Sciences*, 19:12993–12996, 2001.

[479] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Guyon et al. [181], pages 5998–6008.

[480] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Evaluating fidelity of explainable methods for predictive process analytics. In Selmin Nurcan and Axel Korthaus, editors, *Intelligent Information Systems - CAiSE Forum 2021, Melbourne, VIC, Australia, June 28 - July 2, 2021, Proceedings*, volume 424 of *Lecture Notes in Business Information Processing*, pages 64–72. Springer, 2021.

[481] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Evaluating stability of post-hoc explanations for business process predictions. In Hakim Hacid, Odej Kao, Massimo Mecella, Naouel Moha, and Hye-young Paik, editors, *Service-Oriented Computing - 19th International Conference, ICSOC 2021, Virtual Event, November 22-25, 2021, Proceedings*, volume 13121 of *Lecture Notes in Computer Science*, pages 49–64. Springer, 2021.

[482] Ilya Verenich, Marlon Dumas, Marcello La Rosa, and Hoang Nguyen. Predicting process performance: A white-box approach based on process models. *J. Softw. Evol. Process.*, 31(6), 2019.

[483] Maxim Vidgof, Stefan Bachhofner, and Jan Mendling. Large language models for business process management: Opportunities and challenges. In Chiara Di Francescomarino, Andrea Burattin, Christian Janiesch, and Shazia W. Sadiq, editors, *Business Process Management Forum - BPM 2023 Forum, Utrecht, The Netherlands, September 11-15, 2023, Proceedings*, volume 490 of *Lecture Notes in Business Information Processing*, pages 107–123. Springer, 2023.

[484] Marco Virgolin and Saverio Fracaros. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artif. Intell.*, 316:103840, 2023.

[485] Kate Vredenburgh. The right to explanation. *Journal of Political Philosophy*, 30(2):209–229, 2021.

[486] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[487] Toni Waefler and Ute Schmid. Explainability is not enough : Requirements for human-ai-partnership in complex socio-technical systems. In *Proceedings of the 2nd European Conference on the Impact of Artificial Intelligence and*

*Robotics (ECIAIR 2020) / ed. by Florinda Matos. Lissabon: ACPIL, 2020, S. 185-194. - ISBN 9781912764747*, pages 185–194. Otto-Friedrich-Universität, Bamberg, 2021. Jahr der Erstpublikation: 2020.

[488] Michael R Waldmann. Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1):53, 2000.

[489] Douglas Walton. *Informal logic: A pragmatic approach*. Cambridge University Press, 2008.

[490] Douglas N Walton. Begging the question as a pragmatic fallacy. *Synthese*, 100(1):95–131, 1994.

[491] Christian Warmuth and Henrik Leopold. On the potential of textual data for explainable predictive process monitoring. In Marco Montali, Arik Senderovich, and Matthias Weidlich, editors, *Process Mining Workshops - ICPM 2022 International Workshops, Bozen-Bolzano, Italy, October 23-28, 2022, Revised Selected Papers*, volume 468 of *Lecture Notes in Business Information Processing*, pages 190–202. Springer, 2022.

[492] Alexander Warnecke, Daniel Arp, Christian Wressnegger, and Konrad Rieck. Evaluating explanation methods for deep learning in security. In *IEEE European Symposium on Security and Privacy, EuroS&P 2020, Genoa, Italy, September 7-11, 2020*, pages 158–174. IEEE, 2020.

[493] David S. Watson. The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds Mach.*, 29(3):417–440, 2019.

[494] Jochen De Weerdt and Artem Polyvyanyy, editors. *Intelligent Information Systems - CAiSE Forum 2022, Leuven, Belgium, June 6-10, 2022, Proceedings*, volume 452 of *Lecture Notes in Business Information Processing*. Springer, 2022.

[495] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness,*

*Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery.

[496] Sven Weinzierl, Sandra Zilker, Jens Brunk, Kate Revoredo, Martin Matzner, and J¨org Becker. XNAP: making lstm-based next activity predictions explainable by using LRP. In Adela del-Río-Ortega, Henrik Leopold, and Flávia Maria Santoro, editors, *Business Process Management Workshops - BPM 2020 International Workshops, Seville, Spain, September 13-18, 2020, Revised Selected Papers*, volume 397 of *Lecture Notes in Business Information Processing*, pages 129–141. Springer, 2020.

[497] Daniel J Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler, and Gerald Jay Sussman. Information accountability. *Communications of the ACM*, 51(6):82–87, 2008.

[498] Bemali Wickramanayake, Zhipeng He, Chun Ouyang, Catarina Moreira, Yue Xu, and Renuka Sindhgatta. Building interpretable models for business process prediction using shared and specialised attention mechanisms. *Knowl. Based Syst.*, 248:108773, 2022.

[499] Bemali Wickramanayake, Chun Ouyang, Catarina Moreira, and Yue Xu. Generating purpose-driven explanations: The case of process predictive model inspection. In Weerdt and Polyvyanyy [494], pages 120–129.

[500] Boris Wiegand, Dietrich Klakow, and Jilles Vreeken. Discovering interpretable data-to-sequence generators. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 4237–4244. AAAI Press, 2022.

[501] Daniel A Wilkenfeld and Tania Lombrozo. Inference to the best explanation (ibe) versus explaining for the best inference (ebi). *Science & Education*, 24:1059–1077, 2015.

[502] Robert A Wilson and Frank Keil. The shadows and shallows of explanation. *Minds and machines*, 8:137–159, 1998.

[503] Walt Woods, Jack Chen, and Christof Teuscher. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nat. Mach. Intell.*, 1(11):508–516, 2019.

[504] Haifeng Xu, Jianfei Pang, Weiliang Zhang, Xuemeng Li, Mei Li, and Dongsheng Zhao. Predicting recurrence for patients with ischemic cerebrovascular events based on process discovery and transfer learning. *IEEE J. Biomed. Health Informatics*, 25(7):2445–2453, 2021.

[505] J Frank Yates, Ju-Whei Lee, and Julie GG Bush. General knowledge overconfidence: cross-national variations, response style, and "reality". *Organizational behavior and human decision processes*, 70(2):87–94, 1997.

[506] Hangu Yeo, Elahe Khorasani, Vadim Sheinin, Irene Manotas, Ngoc Phuoc An Vo, Octavian Popescu, and Petros Zerfos. Natural language interface for process mining queries in healthcare. In Shusaku Tsumoto, Yukio Ohsawa, Lei Chen, Dirk Van den Poel, Xiaohua Hu, Yoichi Motomura, Takuya Takagi, Lingfei Wu, Ying Xie, Akihiro Abe, and Vijay Raghavan, editors, *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, pages 4443–4452. IEEE, 2022.

[507] Linda Trinkaus Zagzebski. *Epistemic authority: A theory of trust, authority, and autonomy in belief*. Oxford University Press, 2012.

[508] Francesca Zerbato, Pnina Soffer, and Barbara Weber. Process mining practices: Evidence from interviews. In *20th International Conference, BPM 2022, Münster, Germany, September 11–16, 2022, Proceedings*, page 268–285, Berlin, Heidelberg, 2022. Springer-Verlag.

[509] Chiliang Zhang, Zhimou Yang, and Zuochang Ye. Detecting adversarial perturbations with saliency. *CoRR*, abs/1803.08773, 2018.

[510] Hengtong Zhang, Jing Gao, and Lu Su. Data poisoning attacks against outcome interpretations of predictive models. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao, editors, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2165–2173. ACM, 2021.

[511] Jiehuang Zhang and Han Yu. A methodological framework for facilitating explainable AI design. In Gabriele Meiselwitz, editor, *Social Computing and Social Media: Design, User Experience and Impact - 14th International Conference, SCSM 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 - July 1, 2022, Proceedings, Part I*, volume

13315 of *Lecture Notes in Computer Science*, pages 437–446, Online, 2022. Springer.

[512] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In Srdjan Capkun and Franziska Roesner, editors, *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, pages 1659–1676. USENIX Association, 2020.

[513] Tongyu Zhou, Haoyu Sheng, and Iris Howley. Assessing post-hoc explainability of the BKT algorithm. In Markham et al. [306], pages 407–413.

# List of Figures

# List of Tables



This thesis investigates Explainable Artificial Intelligence (XAI) in the context of process mining. The complexity of systems often constrains stakeholder engagement. An interdisciplinary XAI framework is developed, incorporating AI ethics, human-computer interaction, and process mining. It proposes a multi-layered approach and a framework for mitigating risks. The thesis emphasizes adapting explanations to user preferences and mitigating biases in collaborative mining. The findings highlight the interactions between explainability, process mining, and humans for responsible AI integration. The thesis provides recommendations for implementing ethical and comprehensible explanations.