Managing the tension between opposing effects of explainability of artificial intelligence: a contingency theory perspective

Babak Abedin
Macquarie Business School, Macquarie University, Sydney, Australia

Received 9 June 2020
Revised 26 November 2020
6 February 2021
18 May 2021
Accepted 19 May 2021

Abstract

Purpose – Research into the interpretability and explainability of data analytics and artificial intelligence (AI) systems is on the rise. However, most recent studies either solely promote the benefits of explainability or criticize it due to its counterproductive effects. This study addresses this polarized space and aims to identify opposing effects of the explainability of AI and the tensions between them and propose how to manage this tension to optimize AI system performance and trustworthiness.

Design/methodology/approach – The author systematically reviews the literature and synthesizes it using a contingency theory lens to develop a framework for managing the opposing effects of AI explainability.

Findings – The author finds five opposing effects of explainability: comprehensibility, conduct, confidentiality, completeness and confidence in AI (5Cs). The author also proposes six perspectives on managing the tensions between the 5Cs: pragmatism in explanation, contextualization of the explanation, cohabitation of human agency and AI agency, metrics and standardization, regulatory and ethical principles, and other emerging solutions (i.e. AI enveloping, blockchain and AI fuzzy systems).

Research limitations/implications – As in other systematic literature review studies, the results are limited by the content of the selected papers.

Practical implications – The findings show how AI owners and developers can manage tensions between profitability, prediction accuracy and system performance via visibility, accountability and maintaining the "social goodness" of AI. The results guide practitioners in developing metrics and standards for AI explainability, with the context of AI operation as the focus.

Originality/value – This study addresses polarized beliefs amongst scholars and practitioners about the benefits of AI explainability versus its counterproductive effects. It poses that there is no single best way to maximize AI explainability. Instead, the co-existence of enabling and constraining effects must be managed.

Keywords Contingency theory, Systematic literature review, Explainable artificial intelligence, Interpretable analytics, Mitigating strategies, Opposing effects

Paper type Research paper

1. Introduction

Artificial intelligence (AI) offers enormously rewarding opportunities, along with new challenges that need to be identified and handled successfully to utilize AI's advantages and minimize its downsides (Rai, 2020; Abedin *et al.*, 2020; Dwivedi *et al.*, 2019; Beydoun *et al.*, 2019). Humans usually lack understanding about how AI systems produce online behavior analytics or display particular behaviors. This can undermine users' trust in the system and lead to system underutilization, particularly when the effects on individuals are severe, such as when algorithmic prejudices disadvantage certain communities. Humans are reluctant to adopt systems that are not directly understandable and trustworthy (Goodman and Flaxman, 2017).

Internet Research
Vol. 32 No. 2, 2022
pp. 425-453
Emerald Publishing Limited
1066-2243
DOI 10.1108/INTR-05-2020-0300

In response, the demand for ethical AI, and particularly interpretable and explainable AI—which aims to make machines' actions more comprehensible to humans—has risen dramatically in recent years (Bellotti and Edwards, 2001).

In AI, explainability usually refers to the degree to which a learning model's internal dynamics can be described in human words (Ntoutsi *et al.*, 2020). Most scholars seek to improve the understandability, transparency and interpretability of AI systems to validate their decision process and build trust in the model and its predictive and prescriptive outputs (Páez, 2019). The need for greater explainability has been recognized by tech giants such as Google, Facebook, Microsoft, Amazon and IBM, which partner on an open platform called Partnership on AI (https://www.partnershiponai.org/) to facilitate public discussions and to improve people's understanding of AI and its consequences. In parallel, the United States' (US) Defense Advanced Research Projects Agency (DARPA) has launched a series of explainable AI projects (Gunning and Aha, 2019), and the European Union introduced its directive on the General Data Protection Regulation (GDPR) (Felzmann *et al.*, 2019).

Much research has already been done to advance explainability algorithms and study and promote explainable AI systems (Adadi and Berrada, 2018; Biran and Cotton, 2017; Doshi-Velez and Kim, 2017; Nassar *et al.*, 2020). However, less has been done to examine the downsides and unexpected consequences of explainability in AI (Robbins, 2019a, b). Exploring explainability and transparency principles from the perspective of machines outcomes raises many questions, such as what explanation means to different audiences, whether it is always possible or feasible for the machine to explain itself, and whether an explainable mechanism would always lead to harmless outcomes for humans or organizations. For instance, Arrieta *et al.* (2020) find that the explainability of AI systems can have unwanted consequences, such as breaches of confidentiality. Gunning *et al.* (2019) point out that explanation may not be important for some AI applications. The explanatory focus can sometimes be inappropriate, too complex to accomplish and even needless. Robbins (2019b) argues that the principle of explainability is misdirected because the key focus should be on making an AI system accountable.

Despite an overall agreement on making AI systems safer and more useful to humans' life and work, there are polarized perspectives in the literature on whether, when and to what extent AI needs to be explainable to the human user. There seems to be a natural tension between what AI systems are expected to do (e.g. achieve high predictive accuracy) and their level of visibility and explainability, because often high-performing systems are the least explainable, and the most explainable (e.g. regressions) are the least precise (Gunning and Aha, 2019). D'Acquisto (2020) argues that explainability can be a complex problem in autonomous machines since machines are geared first to the principles of formal logic and then (possibly) to ethical or legal principles. This complexity often needs in-depth reflection to prevent conflicts between the legal and ethical principles and the formal logic upon which the AI system is based to serve organizations and/or human users.

In response to the issues outlined above, this paper aims to make a valuable contribution to the literature by improving our understanding of the tensions between opposing effects of explainability in AI—an increasingly important challenge for the future development and dissemination of AI systems. We use contingency theory to propose that there is no best single way to maximize AI explainability; rather, contingent environmental, organizational and individual factors need to be considered in a balanced assessment of its use and enforcement. Such understanding is essential for building and maintaining the trust, approval, integrity and compliance of AI systems from the perspective of developers, administrators, policymakers, users and other AI systems stakeholders (Lecue, 2019). This study, therefore, pursues two objectives in its systematic review of the literature. Firstly, it characterizes and examines the tensions between Comprehensibility, Conduct, Confidentiality, Completeness and Confidence in AI—the 5Cs of opposing effects of

explainability of AI. Secondly, the study offers five perspectives about how organizations can manage these opposing effects and make use of explainability while dealing with the tension between benefits and unwanted or unexpected consequences.

The paper initially provides an overview of what explainability of AI entails, followed by what its opposing effects mean and why they are on the rise. Next, we discuss the methodology for conducting the review, then discuss tensions between the 5Cs. The paper then proposes five perspectives for managing these effects and concludes with a discussion of findings and an agenda for future research.

2. Background

2.1 Explainability of AI: what and why?

Historically, explanations of AI emerged first in the context of rule-based expert systems and were viewed as elements of designing a system capable of producing drill-down outputs (Biran and Cotton, 2017; Gregor and Benbasat, 1999). The rise of data analytics and machine learning in various fields (Beydoun *et al.*, 2019; Abedin *et al.*, 2020) has increased the need for universal methods and practices for examining and verifying the structure and intent of AI systems. In the absence of cohesive theories or principles for AI explainability or interpretability (Arrieta *et al.*, 2020), research to date has predominantly blended approaches drawn from various disciplines. However, there appears to be a general understanding of the intent of explainability of AI as techniques that enable humans to observe how an AI system makes decisions, generates outputs and performs its actions (Rai, 2020).

Explainability, interpretability and transparency have often been used interchangeably in the AI literature, although they sometimes refer to different meanings (Ntoutsi *et al.*, 2020). This confusion has been compounded by the substantial increase in recent years in research in various disciplines, leading to the emergence of keywords such as explainable AI (XAI) (Adadi and Berrada, 2018), explicable AI (Robbins, 2019b) and black box explanation (Guidotti *et al.*, 2018). While some scholars have distinguished between explainability, transparency, interpretability and understandability (e.g. Arrieta *et al.*, 2020; Rai, 2020), these notions are interconnected and aim to address ethics, privacy, bias and fairness in AI (Bertino *et al.*, 2019). They all cover prospective elements (awareness of the collection of data before it occurs) and retrospective elements (revisiting how and why decisions were made) (Felzmann *et al.*, 2019).

Adadi and Berrada (2018) present four reasons why explainability of AI systems is needed:

(1) Explain to justify—reasons for the system's outputs and recommendations,

(2) Explain to control—for better visibility over vulnerabilities and unknowns, and effective response to manage them,

(3) Explain to improve—to easily improve the system by understanding how specific outputs are produced and

(4) Explain to discover—to learn new facts and gain new insights.

Further, stakeholders may need different explanations. (Arrieta *et al.*, 2020) distinguishes five key stakeholders of AI systems:

(1) Managers and executive board members, who examine regulatory compliance,

(2) Regulatory entities, which endorse model compliance with the law,

(3) Expert users of the model (e.g. medical practitioners, insurance professionals), who rely on it for domain-specific knowledge,

(4) People affected by the model, who need to understand how the system affects them,

(5) Data scientists and product owners, who research and develop the model.

Production of explanations that underlie AI systems' behavior is dependent on the type of algorithms they use. Some algorithms yield inherently interpretable models (e.g. regressions, decision trees). In contrast, deep learning and neural networks algorithms, which have complex structural and learning mechanisms, generate models that are inherently difficult to interpret to individual users (Rai, 2020). Increased model complexity means greater efforts must be made to explain it, which can lead to the realization of explainability benefits but simultaneously give rise to unwanted or maleficent effects on one or more stakeholders.

2.2 Opposing effects of explainability

Past research on the transparency and explainability of AI has criticized overcomplication of the system that reduces performance or development time or has exaggerated its benefits and potentials for individual end-users. There remains a need to investigate further the unwanted or unexpected consequences that may emerge as a result of the AI system explaining itself (Gunning *et al.*, 2019; Kroll, 2018). Such research would allow better utilization of the benefits of explainability, while managing its consequent—often opposing—effects.

Recent research shows that explanations may not always be needed for particular AI applications. At the same time, other AI studies present cases in which the emphasis on explanation is misplaced, too challenging to accomplish and unnecessary (Robbins, 2019b). For example, in the legal literature, research shows that disclosure of source code or transparency about how the system operates is neither essential to establish pertinent facts about compliance nor enough to support a regulatory audit of governance practices (Kroll, 2018). This becomes even more important when explainability practices are objectionable to those who profit from the system. Furthermore, explainability is sometimes undesirable because comprehensive insights about an AI system can lead to adverse outcomes like gaming or exploiting computer systems (Kroll, 2018).

Smith and Lewis (2011) offer a managerial perspective to conceptualize and identify tensions between opposing effects in organizing resources. They propose that such effects can be described as contradictory yet interrelated effects that are concurrently present and occur over time. In recognizing opposing effects, two components are noteworthy: tensions that are inconsistent and absurd when juxtaposed and responses that embrace the tensions (Lewis, 2000; Schad *et al.*, 2016). Smith and Lewis stress that just identifying tensions is not enough—it is equally important to postulate strategies for managing them.

Conceptualization of tensions between opposing effects of organizing resources has been used as a device to expose the dilemmas that organizations face in their digitalization practices. For instance, scholars have identified the paradoxical tensions in online knowledge production (Majchrzak *et al.*, 2013) and online privacy (Chen *et al.*, 2019) as well as in information technology (IT) governance and virtual teams in well-bounded organizational contexts (Dubé and Robey, 2009). In the AI context, Felzmann *et al.* (2019) draw on human-computer interaction (HCI) literature to argue that from a practical and user-focused perspective, there is no clear use case for when and in what circumstances intelligent systems need to become more explainable. For instance, from a practitioner and industry point of view, the return on explainable AI investment is not well known and could be small. Explanations occur and make sense in a context that includes the tasks, capabilities and expectations of the user of the AI system (Bellotti and Edwards, 2001). Thus, amongst other things, scholars increasingly stress the need to subject interpretations of transparency and explainability in AI to the domain in which the system operates (Lecue, 2019).

2.3 Theoretical underpinning: contingency theory

2.3.1 Overview of theoretical perspectives in AI explainability. Despite large and valuable bodies of research in philosophy, psychology and the cognitive science of human decision-making, most work in explainable AI relies on the researchers' intuition of what constitutes a "good" explanation (Miller, 2019). Yet, drawing from social science and other relevant theories is key to making explainable AI useable (Miller, 2017). While most studies, particularly technical papers in AI and data analytics, lack theoretical social and psychological implications and inputs (e.g. Felzmann *et al.*, 2019), scholars in other fields are increasingly making contributions that reference the theoretical behavioral underpinning of AI explainability.

For instance, Wang *et al.* (2019) draw from the fields of philosophy and psychology to offer some perspectives on designing theory-driven user-centric explainable AI. Hoffman and Klein (2017) discuss several theoretical foundations of what explanation entails and how people understand explanations, discuss causation versus causal reasoning and demonstrate theoretical implications of the close relationship between explanation and abductive inference. However, their theoretical discussion does not outline the implications of XAI outcomes for stakeholders and their reasoning goals. Guzman and Lewis (2020) address the disconnect between communication theory and emerging technologies and argue that the interactions between AI systems and humans do not fit neatly into traditional communication theory paradigms. They draw on a human-machine communication framework and call for more research into (1) the functional aspects through which users make sense of AI systems as communicators, (2) the relational dynamics through which users associate with these systems and, in turn, relate to themselves and peers and (3) the metaphysical implications called up by unclear boundaries between humans, AI systems and communication. Mohamed *et al.* (2020) explore the role of postcolonial and decolonial theories in understanding AI and establishing ethical principles for protecting vulnerable peoples. They argue that a decolonial critical approach can help AI communities develop insights and tactics to promote and guide transparent and accountable AI systems development.

However, much more needs to be done in theorizing human–AI interactions, particularly in understanding the contradictory effects of AI system explainability and transparency. While business and information systems scholars have been contemplating the theoretical foundations of AI systems since the late 1990s (e.g. Gregor and Benbasat, 1999), their primary focus has been on the technical evolution of AI. Future researchers have been invited to challenge existing theories and utilize their potential for addressing complex questions in AI and data analytics (Berente *et al.*, 2019). In line with this invitation, we draw on contingency theory as a theoretical perspective that acknowledges multiple pathways and discuss how this perspective can encourage future research into AI explainability while considering its unwanted effects. As outlined in the next section, we chose this theory over other perspectives since it explicitly assumes there is no best solution to a problem. This fits the complex problem of explainability in AI systems as the theory suggests taking the circumstances of the problem into consideration and encourages problem solvers to engage with the context in finding a suitable solution.

2.3.2 Contingency theory. Contingency theory is a management theory or model that originated in organizational theory about leadership effectiveness (Feidler, 1964). It has become increasingly accepted because it opposes traditional management theory's contention that there is one best way of doing things (Csaszar and Ostler, 2020). Contingency theory offers two key principles for task-performance fitness: that there is no best way to do things or manage organizations and that a task may be conducted differently in different organizations depending on environmental and contextual factors (Galbraith, 1973). It emphasizes the uncertainty of the environment in which the organization operates and describes effectiveness as organizations coming to terms with their environment(s) to

achieve desirable performance for given tasks (Tosi and Slocum, 1984). While the theory does not explicitly identify the underlying concepts of effectiveness, Tosi and Slocum (1984) point out that a balance between profitability (or service delivery for non-profit), satisfaction and social responsibility is key.

Contingency theory has been widely used in other disciplines, including information systems (IS) and IT management literature. While in the IS literature, it was initially used to study systems design and implementation, it was later applied to studying the influence of the environment and other variables on task-performance fitness (Reinking, 2012). Contingency theory suggests there is no single or best way an IS can be used in all situations; rather, contingent factors (i.e. those that cannot be influenced by the organization) need to be considered in engagement with users and other stakeholders to effectively design and use information systems (Shao *et al.*, 2016). Some scholars have criticized early IS research for ill-defined performance expectations and recommended that the theory be used in a less deterministic manner and focus on particular contexts and applications (Weill and Olson, 1989).

This paper uses contingency theory as a theoretical lens for guiding organizations and decision-makers to manage AI explainability effects. The theory's focus on the contexts in which systems operate has been employed in earlier research to subject the explainability of AI to its domain (Lecue, 2019; Bellotti and Edwards, 2001). We draw on the four key principles of contingency theory (Feidler, 1964; Weill and Olson, 1989) to develop perspectives for managing the opposing effects of AI explainability. The four principles are: there is no universal or best way of doing things, the design of organizations and subsystems need to fit their environment, effective organizations need to fit their environments as well as their subsystems, and the needs and goals of an organization are better satisfied when its management style fits both the task (i.e. AI explanation) and the nature of the work (i.e. opposing effects) within the environment (i.e. the context in which the AI system operates).

3. Review method

3.1 Objectives

Research activity and interest in AI, particularly explainability in AI, have grown rapidly in the last few years. Many scholars have synthesized the diverse methods and sometimes contradictory findings of survey and review papers in attempts to guide future research in this domain. Some surveys (e.g. Dwivedi *et al.*, 2019; Kaplan and Haenlein, 2019) provide overviews of AI applications and methods, whereas others (e.g. Adadi and Berrada, 2018; Arrieta *et al.*, 2020) provide useful entry points for scholars and practitioners to explore key elements of the young and rapidly growing body of knowledge related to AI transparency and explainability. The latter authors survey the literature regarding various dimensions and benefits of AI explainability, discuss trends surrounding its sphere, and present key research opportunities in the near term. Similarly, Biran and Cotton (2017) present recent advances on explainability in machine learning models, and Preece (2018) surveys the history of explainability of AI systems as well as its technical challenges, noting that the central challenges of explainability are far from new and arguing that earlier research on explainability of rule-based expert systems offers ideas for making progress toward better understanding of AI today. Guidotti *et al.* (2018) review black box models and create a taxonomy of interpretability practices and explainability methods in the literature.

All previous reviews focus predominantly on the potential and benefits of explainability of AI and its technical challenges, with less attention to presenting an unbiased and comprehensive view of conflicting effects of explainability in different contexts and for different applications and/or stakeholders. While extant literature on transparency, interpretability and explainability of AI occasionally reports on potential unexpected

outcomes of explainability methods and practices, such findings are dispersed, lack detail and are not connected to the bigger picture of the tension between the opposing effects of AI explainability. This gap in reviews to date motivated our attempt to provide a comprehensive examination of the tensions between the opposing effects of AI explainability and the strategies used to manage them.

3.2 Review procedure

We adopted the four-stage guidelines for systematic reviews outlined by Kitchenham (2004) (see Figure 1), which were applied in several previous studies (Erfani and Abedin, 2018; Priharsari *et al.*, 2020).

We used Scopus to identify resources, referring to Rai (2020) and Arrieta *et al.* (2020) for search terms that would reflect explainability in AI. To broaden our search, we incorporated explainability and other terms that might have been used interchangeably in the literature. Thus, we used a combination of (Interpret* OR Understand* OR Explain* OR Transparen*) AND (Artificial Intelligence OR AI). The search was conducted in March 2020 and produced 50,496 results.

Next, we applied the inclusion and exclusion criteria to select studies. We searched the titles, abstracts and keywords of articles from the Information Systems, Computer Science, and Business and Management disciplines published from 2010 to March 2020. We limited our search to this period because contemporary research on AI interpretability and explainability is a recent phenomenon (Adadi and Berrada, 2018; Arrieta *et al.*, 2020). We excluded non-English papers and limited our results to journal articles to ensure the quality of publications. This procedure identified 4,993 articles; we assessed their titles to filter out irrelevant studies, leaving 398. In the third stage, we examined abstracts to identify articles that address transparency, explainability, interpretability, or understandability of AI. This resulted in 31 articles for the final analysis.

Wolfswinkel *et al.* (2013) suggest that forward and backward citation searches should be undertaken to avoid excluding key papers in the field; in contrast, MacDonell *et al.* (2010) suggest a simple scan of references. We used the former approach because the initial primary studies were adequate (31 papers) and realistic given our time constraints. We conducted a backward search by scanning reference lists and used Google Scholar to identify additional related papers (forward search); this resulted in eight additional articles. Hence, the combined result of both the backward and forward searches was 39 articles. Appendix lists and gives brief descriptions of the selected papers.

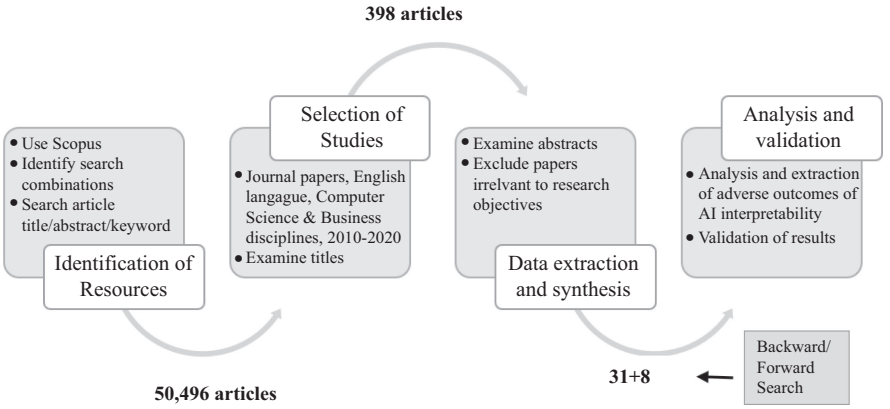

Figure 1.
The steps of the systematic literature review

The screening was conducted independently by the author and a research assistant, and Cohen's kappa was calculated to examine the reliability of the selection (Kitchenham, 2007). Manifestations of reliability were: stability (the process is unchanging over time), reproducibility (replicability) and accuracy (the process conforms to its specification) (Krippendorff, 1989). Screening the title, abstract and body of the text produced a Cohen's kappa value above 0.4 (0.6 for title screening, 0.7 for abstract screening and 0.6 for body text screening), which—according to De Wever *et al.* (2006)—is an acceptable level and confirms the stability and accuracy of the selection. Disagreements in the selection were resolved by combining the first and second screening results. In the last selection stage, disagreements were resolved by reading the body of the text for a third time and reaching consensus.

3.3 Contextual data

Over 80% of the selected papers were published in 2018 or earlier. The trend reflects a considerable rise in interest in explainability and transparency of AI, which unsurprisingly coincides with the growing uptake of AI in fields such as medicine, transportation and finance (Dwivedi *et al.*, 2019; Kaplan and Haenlein, 2019). Importantly, despite a rich tradition in disciplines such as IS and HCI of theorizing and studying the human and technical elements of designing and managing information systems (Dwivedi *et al.*, 2019) , much previous research on understanding and modeling human–AI interactions have been conducted by AI engineers. This is shown because most of the selected articles are in computer science-related journals; they describe AI engineers' and computer scientists' explorations of options for explainability and development of algorithms to make black-box AI more transparent and understandable for the human user. While interest from management and business scholars in AI explainability and transparency has developed relatively slowly, the pace is accelerating as more scholars realize the implications of AI for organizations and individuals. Thus, while sometimes solutions led by AI engineers can produce innovative results and high-speed speed computations (Felzmann *et al.*, 2019), to date, the literature has underappreciated or neglected human, social and organizational perspectives on the complex interactions between humans and AI systems (Harper, 2019).

4. Findings

4.1 Opposing effects of explainability of AI

We used the following criteria, as per Smith and Lewis (2011), to look for tensions between opposing effects of explainability of AI in the selected articles: effects that are interrelated but compete against each other, exist concurrently and can continue, exist within a unified whole (i.e. AI system) and are seemingly logical in isolation but inconsistent when contrasted. As Table 1 summarizes, this led to uncovering five opposing effects, which we call the 5Cs: comprehensibility, conduct, confidentiality, completeness and confidence in AI. We then scanned the selected papers for mitigation strategies, which revealed five perspectives on managing the tension between the 5Cs.

4.1.1 Comprehensibility. Although, articles on explainability in AI have been abundant in the academic and business press in recent years, it is often surprisingly difficult to comprehend what the term explainability or its alternative terms mean in the context of AI. This ambiguity in the definition is inherited from the lack of clarity about what AI itself means; for instance, Kaplan and Haenlein (2019) describe interpretations of AI as being "as white as snow, as red as blood, as black as ebony" (p. 17). The absence of a precise definition has fueled conflict over expectations of explainability (Berkelaar, 2014). Miller (2019) stresses that despite considerable research in philosophy, psychology and cognitive science into how people define, generate, select, evaluate and present explanations, most work on explainable AI uses only the researchers' intuition about what constitutes a "good" explanation.

Opposing effects	Description	Key sources
Comprehensibility	Understandability of what explainability of AI means to different stakeholders, especially laypeople	Adadi and Berrada (2018), Arrieta *et al.* (2020), De Graaf and Malle (2017), Harper (2019), Holzinger *et al.* (2019), Miller (2019), Ntoutsi *et al.* (2020), Páez (2019)
Conduct	The innovativeness and performance of AI systems and their explainability	D'Acquisto (2020), Diez-Olivan *et al.* (2019), Gunning and Aha (2019), Gunning *et al.* (2019), Robbins (2019a), Silver *et al.* (2016)
Confidentiality	Confidentiality, security and safety risks due to the explainability of AI systems	Arrieta *et al.* (2020), Holzinger *et al.* (2019), Tóth (2019)
Completeness	Consequences due to AI logic that cannot be proved or explained	Arrieta *et al.* (2020), D'Acquisto (2020), Silver *et al.* (2016)
Confidence in AI	Overconfidence in AI outcomes vs. mistrust of AI due to its explanation	Bertino *et al.* (2019), Harper (2019), Mittelstadt *et al.* (2016), Pieters (2011)

Table 1.
The 5Cs—opposing effects of AI explainability

Contemporary explainability researchers have started to ask the fundamental question *What is an explanation?* (Adadi and Berrada, 2018). Scholars have begun to challenge or reject usage of the terms "explanation" or "transparency", stressing that while they may be useful for developers or domain expert, they might be not comprehensible for laypeople, especially because (for instance) philosophers have an entirely different interpretation of what explanations are (Ntoutsi *et al.*, 2020). What is implied by explanation has to do with how AI systems are "intelligent" in ways that are similar to human intelligence and thus need to compensate for their intelligence in the same way as a person; they need to justify their actions as, for example, a child might when rebuked (Harper, 2019). Páez (2019) argues that the quest for explainability in AI should be rewritten in terms of the wider goal of providing a realistic and naturalistic interpretation. Intuitively, the objective of presenting an interpretation of a concept or judgment is to make it comprehensible to its stakeholders. Without a prior understanding of what it means to suggest that an agent recognizes a concept or a judgment, explanatory strategies lack a well-defined objective.

There appear to be two schools of thoughts about defining explainability in AI, which overlap and, to some extent, contradict. One represents scholars who argue for unbiased explanations by machines in a way that idealizes a precise explanation that is comprehensible by humans. For instance, Guidotti *et al.* (2018) associate the notion of explanation to an interface between humans and a decision-maker, simultaneously, both an accurate proxy of the decision-maker and comprehensible to humans. Arrieta *et al.* (2020) distinguish between explainability, understandability, transparency and interpretability and define explainability in AI as "given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand" (p. 84). In this definition, they note that machines may have different audiences (e.g. users, decision-makers, policymakers) and that a concise reason for AI needs to be provide and customized for each.

On the other hand, Miller (2019) argues that people use particular prejudices and societal perceptions when creating and evaluating explanations that can enhance human experiences with explanatory AI. Since people assign human-like traits to artificial agents, they will expect explanations using the same conceptual framework used for human behaviors (De Graaf and Malle, 2017). While in an ideal sphere, AI and human systems would be equal and correspond with ground truth, this might not occur in the real world due to two problems (Holzinger *et al.*, 2019). Ground truth is often hard or sometimes impossible to define, especially in contexts like medicine; man-made systems and models often look for causal reasons and justifications for understanding fundamental mechanisms. In comparison,

AI systems are based on models that often provide only a probability analysis for further establishing causal argumentations. Thus, some scholars offer an alternative view of explainability and argue that the real aim is rather "explicability" of the result of the process—not the explainability of the process itself (Robbins, 2019b). This call for explicability places emphasis on requiring stakeholders to provide explanations when needed rather than requiring them for every decision. The same approach can apply to AI systems.

4.1.2 Conduct. There is an ongoing and seemingly natural tension between the conduct of AI systems (i.e. performance) and their explainability (Gunning and Aha, 2019; Silver *et al.*, 2016). This is due to considerable research on deep learning and black box models for making faster and more accurate AI systems working—intentionally or unintentionally—against making these systems more explainable (Gunning *et al.*, 2019). This problem requires considering trade-offs, including precision and fidelity, to forge a balance between performance and explainability.

It is often assumed that top-performing AI systems get the lowest marks for explainability, on the basis that approximation involves complexity because models require vast amounts of data to be collected and complex multi-variable models to be used for interpretation, learning and analysis (Diez-Olivan *et al.*, 2019). However, more recently, scholars have begun to question this assumption (Robbins, 2019b) and argue that more complex models are not necessarily more accurate. Diez-Olivan *et al.* (2019) claim that the assumption is incorrect when the data are well structured, and features at our disposal are of great quality and value. However, it is true that more complex models are likely to be more flexible than their simpler counterparts and thus suitable for more complicated tasks. Diez-Olivan *et al.* stress that using inappropriate complex predictive models falls into the trap of overcomplicating problems, especially when insufficient data diversity (variance) exists. Thus, the added complexity of the model will only work against the explainability of the system.

D'Acquisto (2020) emphasizes the value of a balanced perspective in the battle between performance and explainability and points out that while a "certain level of transparency" (as opposed to a "certain level of autonomy") of black box AI helps reduce mistrust in the system, the quest for explainability and transparency should not destabilize other principles and logical constraints (p. 899). Unreasonable and unjust AI explainability requirements can be a disincentive for innovation, especially as innovations are often protected by specific intellectual property laws with limited openness to external stakeholders. Setting a regime of AI explainability while stimulating performance and innovation can protect the interests of system owners and domain expert users.

4.1.3 Confidentiality. Holzinger *et al.* (2019) question the assumption that humans are always able to provide an explanation for their decisions and stress that experts are often unable to offer reasons for their decisions due to reasons such as heterogeneous and vast sources of information, as well as confidentiality, safety, security, privacy and ethical reasons. For instance, Arrieta *et al.* (2020) point to adversarial attacks. External parties try to manipulate an algorithm after learning how it operates, as an important security threat as a result of explainability. Attacks on a supervised machine learning classification model, for example, would reveal the minimum changes needed to be applied to the inputs to create a different classification output.

A challenge in dealing with algorithmic decision-making and analytics is that their underlying codes are often trade secrets and thus hidden from the public eye (Tóth, 2019); this maintains their competitive advantage, as well preventing the system from being "played". For some organizations, protecting sensitive and novel algorithmic systems is a top priority, but this work against the system's transparency and explainability to outsiders and makes compliance challenging. The resulting non-explainable system can end in organizations overprotecting their AI machines, leading to dereliction of responsibility and abuse of power.

The contradiction between confidentiality and explainability will be a major challenge as AI use and adoption, as well as competition between AI developers and owners, rise in coming years (Arrieta *et al.*, 2020), even though cutting-edge AI technology offers new promise for algorithmic copyright enforcement (Tóth, 2019). Imagine a domain-specific algorithm that a company has developed over several years of research and investment. The company may consider the knowledge synthesized in the algorithm confidential, and as such, may rightfully find it compromised even by providing basic information about its input and output for explainability purposes. Explaining the system can enable the development of techniques for attacking and confusing the system and more accurate and efficient ways to improve the system's privacy and security (Arrieta *et al.*, 2020).

4.1.4 Completeness. The completeness constraint in AI has been described as "the absence of a possible use of the machine beyond design requirements" (D'Acquisto, 2020, p. 896). Completeness is a by-product of Gödel's first incompleteness theorem (Gödel, 1931), which suggests that logics based on a set of axioms, along with rules of symbolic combinations of statements about those axioms, cannot be proved or explained in formal systems (such as an algorithm). This has major implications for the design and explanation of AI (D'Acquisto, 2020).

Research on AI completeness aspires to ensure machines do not generate or lead to the generation of harmful outputs (Silver *et al.*, 2016). More specifically, the goal of ensuring AI systems act ethically and non-maleficent requires a realistic technical design that stops the system from reaching any known harmful states in which its behavior can be considered dangerous to humans (Arrieta *et al.*, 2020). However, achieving this goal needs to deal with two challenges: (1) AI may progress to a point currently understood as safe, but which may later turn out to be unsafe, for instance, in specific complex models, and (2) there may exist unfamiliar or unobservable conditions that humans can only infer, which are reachable by AI and unsafe to humans—for instance, when an AI system is used in an unprecedented application (Arrieta *et al.*, 2020).

Silver *et al.* (2016) stress that in examining completeness in AI, researchers should consider that humans' cognitive abilities, and thus their ability to process and understand complex AI models, are limited. These limitations can lead to incompleteness in AI systems, which can make AI explainability a partially achievable goal. Silver *et al.* argue that factors such as the competitiveness or unpredictability of the contexts in which AI systems operate or interact with humans can make explainability of the system merely wishful thinking, which leaves us wondering if humans should strive for less ambitious but more practical explainability expectations.

While explanations can help system developers, decision-makers and other stakeholders to understand AI systems better or to challenge their incorrect assumptions, one needs to note that explanations alone do not produce understanding (Kroll, 2018). In order to reduce the chances of unwanted harmful outcomes, the explanation needs to be complemented by other evidence to be believable, cover potential causes of an outcome and engage and be targeted to particular stakeholders (Preece, 2018).

4.1.5 Confidence in AI. Trust has been discussed widely in the technology use literature as a key factor in AI explainability. Similarly, gaining users' and other stakeholders' trust is considered a major pillar of the explainability of AI systems (Páez, 2019; Lecue, 2019). Yet, Mittelstadt *et al.* (2016) show that while explainability of AI systems is important to maintain a trusting relationship with data subjects, a lot of confidence is already placed in some AI systems, especially if the system provider is well known to the user or the system has been in use for a long time. While increased confidence in AI can be associated with increased AI use, overconfidence in AI outputs can lead to de-responsibilization of human stakeholders or a tendency to "hide behind the computer" and assume that AI outputs are correct by default. Thus, it is important to determine the level of trust in AI that would be needed or is even possible and the expected and unexpected consequences of trust in these systems.

Pieters (2011) makes a distinction between trust and confidence and suggests that for trust, unlike confidence, risks are apparent and compared. Pieters argues that explainability makes AI systems more observable, which creates two scenarios: explanation-for-trust, which allows users to compare options by describing them in detail and explanation-for-confidence, which enables them to be confident in using a system without having to consider options. In the former, the AI black box needs to be "opened" to build trust, whereas in the latter, it is often used to give the user confidence in the system's outputs without the need to open it. Felzmann *et al.* (2019) emphasize the contexts in which AI and humans operate and interact and that it is in some conditions that many things or interactions can be considered trustworthy. This, in turn, implies whether explainability is always needed for establishing trust with the user. Miller (2019) highlights that the quest to build trust in AI systems is not just a computational or machine-related problem. As stressed by other scholars (Bertino *et al.*, 2019), the involvement of several other actors (e.g. those who examine or audit compliance of an AI system) needs to be considered in establishing trust and confidence. Sometimes these parties are semi-trusted, and entrusting them with AI codes could give them the power to risk the interests of the AI owner or developer or violate transparency regulations. Harper (2019) argues that building trust between human and AI actors requires scrutiny of complex interactions between and within them, which requires a multidisciplinary engagement between business, IS and HCI scholars and AI engineers. Harper suggests that a lack of such engagement has often led to a less-than-comprehensive design of human–AI interactions by technical engineers. While some of their designs are innovative, they are not good solutions in terms of interface and interactions with humans, damaging trust in the system.

4.2 Perspectives for managing explainability of AI

In the following, we unpack six perspectives that can be used in managing tensions between the opposing effects of explainability in AI. Among them, contextualization of explanation emerged as a pivotal element that could shape, justify, reinforce and make sense of the application of explainability techniques and procedures in a particular domain, as well as the management of the tensions between its opposing effects. Table 2 provides a summary of these findings.

4.2.1 Pragmatism in explainability. Explainability in AI and its effects, consequences and expectations start with its initial intentions and what we mean by it. Earlier, we presented various perspectives on expectations of explainability and their contradictory effects due to the lack of a comprehensive definition. In addressing this, Garibaldi (2019) stresses that if human experts are prone to error and are not expected to achieve 100% performance in all tasks and explain the underlying reasons, then why should we not expect the same from "expert" computer systems? An answer to this question is Robbins' (2019a) argument that the property of requiring explainability and accountability should be assigned to a particular task or output rather than the entity (i.e. AI system) undertaking the task. This invites scholars and practitioners to focus on the contexts of tasks and their potential damage rather than the process by which tasks are undertaken. This makes more sense when AI is employed for low-risk purposes, for which justification is unnecessary. Requiring explainability and accountability from AI would prevent or slow down the realization of the benefits of AI in many low-risk and some high-risk situations (Robbins, 2019b; Waltl and Vogl, 2018). When Google's AlphaGo defeated the world champion Go player Ke Jie, it did not offer any explanations for its moves, and this lack of explanation did not concern or hurt anyone. As such, Robbins (2019a) argues that many AI systems fall into this category.

Based on the arguments outlined above, scholars are increasingly calling for a more pragmatic approach in understanding and interpreting explainability in AI, one that relies on the context, balanced agency between AI and humans, and measurement of explainability. For instance, Felzmann *et al.* (2019) propose to understand explainability relationally,

Perspectives	Description	Key sources
Pragmatism in explainability	Understanding explainability relationally and assessing the trustworthiness of AI in its communication with humans based on contextual and other factors that mediate the value of the explainability of the communication	Felzmann *et al.* (2019), Garibaldi (2019), Miller (2019), Robbins (2019b)
Contextualization of the explanation	Assessing tasks, capabilities and expectations of the AI system based on its context, which is everything that shapes and influences our perceptions, cognition and actions in a particular domain	Bellotti and Edwards (2001), European Commission (2019), Lawless *et al.* (2019), Lecue (2019), Miller (2019)
Cohabitation of human agency and AI agency	Examining how responsibility for AI decisions and consequences should be distributed between humans and the AI system	Adadi and Berrada (2018), Coeckelbergh (2009, 2020), D'Acquisto (2020), Kroll (2018)
Metrics and standardization	Context-specific as well as universal measures that can compare, evaluate and quantify explainability methods and their efficiency, outcomes and impacts	Arrieta *et al.* (2020), Doshi-Velez and Kim (2017), Garibaldi (2019), Gunning and Aha (2019), Gunning *et al.* (2019), Preece (2018)
Regulatory and ethical principles	Ethical and legal frameworks that ensure and promote human safety and autonomy in interactions with AI and maintain and encourage visibility and explainability in the use and propagation of AI systems	Bertino *et al.* (2019), Hacker *et al.* (2020), Robbins (2019b), Stahl and Wright (2018), Tóth (2019)
Other emerging solutions	Innovative and emerging solutions, particularly incorporating AI enveloping, blockchain and fuzzy systems, for improving and managing explainability in AI	Bertino *et al.* (2019), Fernandez *et al.* (2019), Floridi (2011), Garibaldi (2019), Kshetri (2019), Nassar *et al.* (2020), Robbins (2019b)

Table 2.
Perspectives for managing explainability of AI

meaning the interaction between humans and AI is considered communication between technology and its stakeholders. Its trustworthiness is gauged by contextual and other factors that mediate the value of explainability of the communication. In helping to operationalize such a pragmatic understanding of explainability, Miller (2019) proposes four characteristics:

(1) Explanations should be contrastive—responsive to particular counterfactual events. For instance, we do not ask why event P occurred, but we ask why event P happened instead of some event Q;

(2) Explanations should be selected. People rarely expect a full explanation that contains all possible causes of an event. Rather, influenced by certain cognitive biases, they choose one or two causes from all potential causes;

(3) Probabilities probably do not matter. While people appreciate truth and likelihoods, referring to probabilities or statistics in explanation is not as effective as knowing the causes; and

(4) Explanations are social. People understand explanations as part of the information exchanged in a conversation or interaction relative to their beliefs and background.

4.2.2 Contextualization of the explanation. In its recent communication "Artificial Intelligence for Europe" (European Commission, 2018), the European Commission portrays the distinctive characteristics of AI systems in the presence of "a certain degree of autonomy" in decision-making. It stresses that the crucial issue in the design and dissemination of these systems is the context in which they operate. Explanations occur and make sense in a context that includes the tasks, capabilities and expectations of the user of the AI system (Bellotti and Edwards, 2001). Thus, interpretations of transparency and explainability in AI are domain-dependent and should not occur in isolation from the particularity of the domain. However, context is rarely considered in existing research on AI explainability (Lecue, 2019).

Lawless *et al.* (2019) stress the importance of context in managing AI systems and describe context as everything that shapes and influences our perception, cognition and actions in a given environment. They promote the interdependence between humans and AI, which needs to be mastered and properly designed for efficient human–AI teamwork. Miller (2019) builds on the four characteristics above of explanation in AI, emphasizing that in understanding AI outputs, the focus should not just be on associations and causes, but on the context in which the system operates. This study proposes three important points that need to be considered in building truly explainable AI systems: (1) of various potential causes of an event, the explainee usually cares about *a small subset (located in the context)*, (2) the explainer *picks a subset (founded on various criteria)* for offering the explanation and (3) an interaction or argument may occur between *explainer and explainee* about this explanation.

Overall, contextualization of the explanation appears to rely heavily on the interaction between the actors, which are the AI system and the users or stakeholders influenced by the system outputs. Bellotti and Edwards (2001) argue that there exist human elements of context that the machine cannot recognize. Therefore, AI systems cannot be designed or expected to act on our behalf. Instead, the system needs to comply with users and relevant regulations in an efficient and non-obtrusive fashion. These authors put stress on the interaction between humans and AI systems and the need to conceptualize the expectations from each party. They highlight that it is unlikely that AI developers and designers can sufficiently clearly determine human aspects of context to allow the system to act on humans' behalf. Rather, what is needed is a set of design principles to enable humans to understand and be guided about interactions with the AI system and to the reason for themselves how best to proceed.

4.2.3 Cohabitation of human agency and AI agency. The third perspective aims to help improve our understanding of the interactions between human and AI systems and shed some light on responsibility for the outcomes. Research on the explainability and transparency of AI has predominantly focused on technical and algorithmic factors, and relatively little has been done on human factors and interactions between human and AI systems (Adadi and Berrada, 2018; Kroll, 2018). Human factors are equally (if not more) important considerations in understanding AI explainability. For instance, in medicine, while explainability of AI systems is very necessary for applications such as education, research and clinical decision-making, human experts must stay engaged in these processes to understand and to review the AI systems' decision processes and outputs (Holzinger *et al.*, 2019).

Research on the explainability of AI is still unsettled with respect to the agency of AI and how responsibility for AI decisions and consequences should be distributed between humans and AI systems. Extant literature widely assumes that only humans should be regarded as responsible agents (Adadi and Berrada, 2018). The idea that all responsibility can be allocated to humans and not machines assumes that AI systems are not stakeholders and have no interests to defend (D'Acquisto, 2020). This idea suggests that regulations and ethics are only applicable to humans and that it is humans' responsibility to allow or avoid the point of no return of AI autonomy, and therefore humans are the only responsible agents for AI system decisions.

In contrast to the concepts discussed above, scholars are increasingly asking about and arguing for the agency of the AI system. Ågerfalk (2020) posits the idea of digital agency for AI and stresses that information systems like AI are not solely demonstrating an external reality; rather, these systems are dynamic participants in reality which also engage in forming it. Coeckelbergh (2009) points out that AI systems do things that influence us in many ways, and what happens as the result of this influence needs to be discussed in terms of right and wrong, which in turn implies that we can regard AI systems as agents or moral agents. Coeckelbergh argues that if non-humans (natural and artificial) can significantly affect our lives, it is undesirable and unhelpful to exclude them from moral equations. This raises the idea of whether AI systems themselves can be responsible agents (Coeckelbergh, 2020), which in turn engenders the idea of a hybrid approach in deciding the responsibility of AI systems. In this hybrid approach, human agency and AI agency would coexist, and responsibility for the outputs would be distributed across a network of humans and machines, acknowledging that:

> 1) technology shapes human action in a way that goes beyond a merely instrumental role, that (2) the actions of AI-driven technology can be morally relevant. . ., and that (3) advanced AI technologies may give the appearance of being responsible agents. (Coeckelbergh, 2020, p. 2,054)

This discussion highlights that organizations and decision-makers need to make responsibility expectations clear to all stakeholders. Furthermore, societies need to think about strategies that can promote policies that ensure machine outcomes can be controlled and harmful consequences can be assessed and acted upon by a trusted centralized entity (D'Acquisto, 2020). In assessing and managing machine outcomes, individual experiences can be distinguished from group experiences. An explanation of AI outcomes is prioritized for group effects and for safeguarding the public interest.

4.2.4 Metrics and standardization. An important aspect of evaluating AI systems and managing their performance is metrics that measure a model's performance in evaluating a particular aspect of explainability and make sense to stakeholders (Garibaldi, 2019). However, very few of the studies we reviewed provided clear metrics for evaluating or measuring explainability effects or performance. This result is in line with earlier reviews and surveys; for instance, Adadi and Berrada (2018) found very little work on evaluating explainability methods and quantifying their relevance.

Scholars and practitioners are increasingly calling for metrics that can help compare, evaluate and quantify explainability methods (Preece, 2018). The field clearly needs unified concepts for measuring explainability effects and outcomes necessary for enabling and guiding future research on techniques and methods for AI explainability. Metrics would enable clarity in the evaluation of how well a variable would fit with the explainable description. Without such mechanisms, every argument in the literature lacks a stable foundation (Mohseni *et al.*, 2018).

While existing methods lack quantitative and comprehensive scales for measuring various aspects of explainability of AI and comparing explainable techniques (Arrieta *et al.*, 2020; Gunning and Aha, 2019; Gunning *et al.*, 2019), some tools offer useful insights for the development of the next generation of explainability metrics. For example, Mohseni *et al.* (2018) present a multiscale tool for measuring AI explainability that includes the goodness of the model fit, explanation satisfaction indicators, assessment of mental models, indicators of the reliability of computations and explanation trustworthiness. D'Acquisto (2020) promotes value transparency in measuring the usefulness of AI systems by disclosing criteria humans and machines can employ to settle disputes. Poursabzi-Sangdeh *et al.* (2018) propose a general model called Network Dissection, which aims to quantify interpretability in terms of alignment with a set of human-interpretable concepts.

Doshi-Velez and Kim (2017) identify three types of interpretability assessments:

(1) Application-grounded—a user (normally a domain expert) tests the explanation in a particular application domain,

(2) Human-grounded—laypeople, rather than experts, assess the explanation in an experimental format and

(3) Functionally-grounded—explanation is assessed through pre-specified models and with no human subjects engaged.

Mohseni and Ragan (2018) build on the above three assessment types and develop a human-grounded evaluation standard for estimating instant explanations of images and textual data. By associating the explanation results from classification models with the evaluation standard, they test the quality and relevance of local explanations. Bertino *et al.* (2019) offer the standardization of transparency data exchange and transparency score; the former stresses the use of acceptable data formats to share transparency information about the data, while the latter presents the level of transparency that is available for a user about a dataset. Finally, Gunning and Aha (2019) develop an evaluation framework for DAPRA that includes five dimensions: user satisfaction (clarity and utility of the explanation, utility of the explanation), mental model (assessing individual decisions and their strength/weakness), task performance (impact of the explanation on the user's decision/task performance), trust assessment (users' trust and the likelihood of AI use in future) and correctability (opportunities for identification and correction of errors).

4.2.5 Regulatory and ethical principles. Despite the increasing amount of research on the explainability of AI, there remains substantial ambiguity about what elements of explainability need to be incorporated in AI systems, how, and to what extent, and how users can be made aware of their rights in interactions with machines. There is an academic and professional consensus that ethical and legal frameworks can help; governments in Australia (Australian Government, 2019), China (Zhang, 2019), the EU (European Commission, 2019) and elsewhere have developed and promoted principles for explainability and ethics in AI. Major technology companies like Microsoft and Google and the World Economic Forum have drafted AI ethics guidelines under the umbrella of "explicability" (Robbins, 2019a). Many of these principles share the general aim of ensuring that AI supports and does not limit human autonomy; to achieve this aim, we need to understand how humans may be affected by AI and make sure we know how AI will act on our behalf (Floridi *et al.*, 2018).

However, many of the frameworks mentioned above are still under development, and there is uncertainty about the specifics of their underlying ethical and legal principles (Bertino *et al.*, 2019; Robbins, 2019a). Bertino *et al.* (2019) highlight the need for regulatory requirements for data transparency to be enforced by authorized entities like governments, standards bodies, or corporate governing authorities to maintain and encourage visibility and explainability in the use and propagation of AI systems. These requirements need to acknowledge that specific criteria must be applied on a case-by-case basis so that different purposes, context and actors can be considered in making judgments about outcomes and effects of AI use and explainability. Tóth (2019) calls for more non-profit and research organizations to develop their own enforcement algorithms to create more balanced competition and transparency for emerging platforms. This will lead to an increased number of actors engaged in AI development, improve understanding of explainability effects and alleviate some of the pressure caused by trade secrecy and competition between for-profit organizations. Stahl and Wright (2018) suggest the notion of responsible research and innovation (RRI) as a framework for ensuring the acceptability and sustainability of technologies in society. The European Commission uses RRI to expand on six key issues: public engagement, ethics, science education, gender equality, open access, governance

(European Commission, 2019, 2020). Yet, despite RRI's benefits in providing a guide for key elements of explainability of AI systems, it lacks a clear understanding of the effects and impacts of AI systems on relevant stakeholders and how to respond to such effects.

There is debate about the regulations or principles that may help with AI explainability and accountability in the law and ethics literature. Hacker *et al.* (2020) point out that most current legal debate focuses on data protection, and particularly on whether the GDPR implies a right to an explanation of automated decisions. Hacker *et al.* (2020) emphasize that the law and AI are interwoven and often mutually reinforcing. They argue that explainability of the AI system is an imperative legal category, not just in data protection law but in contract and tort law. They propose that contract and tort law are more likely than data protection law to lead to enforceable legal requirements in regard to explainable machine learning models.

Bellotti and Edwards (2001) four general categories for accountability of context-aware systems could be used as a basis for identifying ethical and legal principles for enforcing and guiding the use of explainability in AI systems:

(1) Inform the user of the system about its abilities and possibilities,

(2) Offer users feedback about using the system:

- Feedforward—so that the user understands the consequences if they take a particular action,
- Confirmation—so that the user understands what they have done,

(3) Enforce identity and disclosure, particularly with sharing sensitive data and

(4) Provide users with control over the system, especially when user actions may influence them.

4.2.6 Other emerging solutions. Research into making AI more explainable and transparent is dynamic and evolving. Scholars are continuously looking for new techniques to complement existing methods or innovative approaches to advance the field. In particular, the review of the selected articles revealed a growing interest in using AI enveloping, blockchain and fuzzy systems in improving and managing AI explainability.

Robbins (2019a) presents an alternative narrative to AI transparency and explainability, suggesting that instead of requiring transparency from machines, the focus should limit their power within physical and virtual microenvironments. This ensures machines can perform their functions whilst reducing the risk of harm to humans; in robotics, this is called "envelopment" (Floridi, 2011). Floridi offers the example of dishwashers. These machines are properly enveloped within a certain context to perform their operations, which has made them useful and effective tools; the alternative would be an ineffective humanoid dishwashing robot. Robbins (2019a) builds on the idea of envelopment and argues that AI will be effective if it is successfully constrained within an "envelope" to produce desired outputs within a given limited capacity. The key requirement for AI envelopment is a sound knowledge of its inputs, outputs, functions, training data and boundaries, which is often lacking. Researchers working on AI envelopment continue to look for ways to obtain and incorporate such knowledge into the AI system.

To improve the trustworthiness of explainability, Nassar *et al.* (2020) propose a framework that leverages blockchain features. Their framework uses smart contracts, trusted oracles and decentralized storage to model explainability decisions subject to a consensus between distributed agents, with the assumption that the majority of agents involved are truthful. They pose that blockchain-enabled explainability can offer transparency and visibility, immutability, traceability and nonrepudiation, and smart contracts. Kshetri (2019) argue that AI and blockchain complement each other substantially. The study poses that blockchain can

break the dominance of a few major players in AI because blockchain predictions are immutable, and the reputation of AI providers would depend more on their service quality capabilities rather than providers' size and reputation. Bertino *et al.* (2019) stress that to improve transparency and explainability, unified policies and contracts are needed to implement ethical principles. Blockchain is, therefore, in a good position to encode such policies into smart contracts for enforcement among networks of peers and help with tracking the origins of data as it evolves, which in turn provides a dynamic consensus in the network for handling special situations.

Fuzzy sets were introduced in 1965 by Zadeh (1965) specifically to deal with difficult "classes of objects encountered in the real physical world" which are "imprecisely defined" and yet "play an important role in human thinking". Fernandez *et al.* (2019) present a detailed discussion of evolutionary fuzzy systems and show the synergy between fuzzy systems and evolutionary algorithms in AI. They assert that such synergy can offer a decent trade-off between the accuracy and explainability of AI models. They discuss how evolutionary fuzzy systems can incorporate key features of explainability in machine learning systems for making outputs more understandable to stakeholders. Garibaldi (2019) highlights the inherent uncertainty in real-world knowledge and data and that fuzzy systems are well placed to deal with uncertainty in decision support and knowledge representation and inference. Garibaldi presents indistinguishability as the key factor in the assessment of computerized decision support systems and argues that the technical and non-technical aspects of AI explainability are interconnected and need to be used together and in an understandable fashion when the system offers explanations to humans. Following this, Garibaldi (2019) seeks answers to the question "can AI explain itself?" and proposes using the Turing Test (Turing, 2009) to answer it. This test poses a problem involving an AI system and humans and checks if the user can distinguish between the system and the human object. If the user cannot make the distinction, then AI presents an acceptable level of performance regardless of its accuracy.

5. Discussion and future research

As AI systems become ubiquitous and as scholars and practitioners look for more advanced algorithms to deal with complex problems, concern about the understandability of these black box systems is rising (Robbins, 2019a). Simultaneously, there is growing interest in the broader societal effects of AI systems and establishing good governance to make these systems accountable (Kroll, 2018), which has led to a substantial increase in research into techniques and strategies for improving AI transparency and explainability.

However, as shown above, much past research has neglected the opposing effects of AI explainability and their consequences. The dilemma of how to deal with opposing effects of explainability is not unique to AI: in information security, the constant question for managers is whether it should be sought *ex ante* (before a security incident happens) by investing in technical and organizational measures or ex-post (after an incident happens) by focusing on ad hoc solutions (Odlyzko, 2019). Through a systematic review of studies in IS, business and computer science published in recent years, this study has made a significant contribution to knowledge by revealing the tensions between opposing effects of explainability of AI. Figure 2 shows a high-level framework that proposes perspectives that can be used to manage the opposing effects of AI explainability, in which context is at the core for integrating the management of AI explainability effects. This paper highlights the fundamental impossibility of complete and non-maleficent AI by stressing the tension between opposing effects of AI explainability.

Our findings show that this tension is unlikely to diminish anytime soon; rather, it is expected to intensify as the use and propagation of AI increases over the coming years.

Organizations and policymakers need to assess and implement context-specific strategies to manage explainability and interpretability impacts and efforts. There is a constant tension between societal and organizational expectations of explainability because AI systems' increasing accuracy and performance will often confront perceptions about the impacts of these systems' outcomes on human stakeholders' safety and autonomy.

Context has always been a key element of the IS and technology management literature because it ensures we:

> stay in touch with the practical context in which information systems are used. It also means that we should not black-box technology, and we should not assume that technologies will work the same or be ascribed the same meaning in all contexts. (Ågerfalk, 2020, p. 6)

While acknowledging the importance of context, we propose six perspectives for bringing this tension under control so that explainability and its benefits can be realized without damaging the efficiency and performance of AI systems. We do so by drawing on contingency theory and posing that expectations of and investments in AI explainability should be contingent on the AI's environment. We build on contingency theory's four premises as key assumptions for the conceptualization of the proposed perspectives:

(1) There is no universal or best way for organizations to do things (e.g. AI explainability), meaning that there is no best way or approach for making AI systems explainable,

(2) The design of organizations and subsystems need to fit the environment, highlighting the importance of the context in which AI systems operate and are being designed for,

(3) Effective organizations need to fit their environment as well as their subsystems, stressing the match between explainability expectations of the AI system in light of ethical and legal principles as well as expectations of other human stakeholders (e.g. AI owner and AI user) and other subsystems (e.g. other AI systems); this also highlights AI and human agency and shared responsibilities and

(4) The needs and goals of an organization are better satisfied when its management style fits both the task (i.e. AI explanation) and the nature of the work (i.e. explainability's opposing effects) within the environment (i.e. AI context).

The first perspective is using a more practical and pragmatic form of explainability; current definitions suit AI engineers and domain experts but are hard for a lay human to understand.

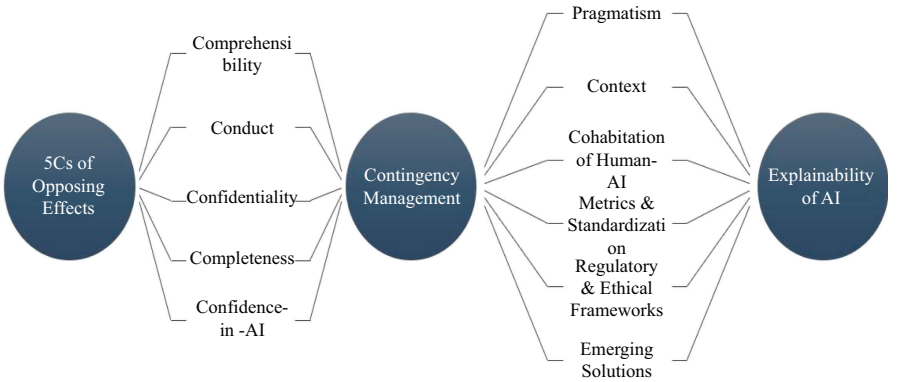

Figure 2.
A framework for managing tensions between the 5Cs (opposing effects of explainability of AI)

A more understandable definition needs to be fitted into the context in which the AI system performs (our findings emphasize the importance of this), and in low-risk situations, there may be no need for an explanation. Contextualization of the explanation appears to rely heavily on the interaction between the actors engaged, so in situations in which an explanation is needed, the explainer (i.e. the AI system) needs to provide the level of explanation that the explainee demands or one that is needed for compliance purposes. While research on domain-specific AI development is on the rise, less has been done to contextualize the explainability of such systems. Future research is needed to specify the explainability needs of domains (e.g. medicine, law, finance) and determine criteria that can guide whether and how explainability is required from such AI systems.

Our next proposed perspective highlights the need to objectify and standardize the current subjective discussions around explainability by incorporating and developing new metrics and standards. Existing research on this topic suggests using general measures such as explainability models' goodness of fit, user satisfaction and model trustworthiness. In particular, Doshi-Velez and Kim's (2017) three types of interpretability assessment (application-grounded, human-grounded, functionality-grounded) seem to be a good starting point for advancing context-specific scales as well as universal measures and standards for determining whether explainability is needed, and if so, how to measure its outcomes and consequences. However, any research on this topic remains premature: no universal or context-specific standard is available to objectively quantify the development of AI systems and stakeholders' expectations of the explainability of the system. This void represents an excellent opportunity for future researchers to develop, test and generalize stakeholder-specific and context-aware metrics for AI explainability needs and outcomes.

The fourth perspective highlights the current interest in framing ethical principles and legal frameworks for guiding AI developers and users to distinguish right from wrong in developing, using and disseminating AI systems. Governments and major enterprises have started introducing such frameworks. Still, all such attempts are in the early stages and require researchers and practitioners to provide input and feedback for further development. Felzmann *et al.* (2019) argue that while an ethical or legal right to explanation can offer useful visibility about an ex-post solution and why a system reaches a decision, it cannot *per se* justify the reasons for such a decision or shield the user from potential harmful consequences. Berkelaar (2014) calls for more research into enhancement of ethical principles in AI explainability and invite more work on protecting users' expectations of an explainable AI by elucidating the social contract between the user and the system, documenting explainability expectations and clarifying the ethical and practical implications of AI decisions for individuals.

The fifth perspective encompasses novel approaches borrowed from robotics, blockchain and fuzzy system literature. Floridi (2011) introduced the notion of envelopment of AI, in which AI systems become constrained in a respecified environment, as an alternative mechanism to explainability. While this may hurt systems' flexibility, it helps to reduce AI incompleteness by restricting its functionalities to boundaries that pose no harm to humans. However, more research is needed to determine the trade-off between gains and losses because of such restrictions. While enveloping AI offers more assurance to users about a system's outcomes and the consequences of its decisions, it may bring down its performance and the "learnability" of the underlying algorithm. The next encouraging stream is blockchain, which promises to improve the transparency and traceability of storage and data and advancing the development and use of unified ethical policies and smart contracts between trusted parties. Bertino *et al.* (2019) invite future researchers to assess how undermining privacy, security, scalability and availability issues in the blockchain would affect explainability mechanisms. They argue that a key goal is to determine the specific extensions (e.g. aggregating/obfuscating/anonymizing) blockchain needs to support the explainability of AI systems. The final element in this category, fuzzy systems, has opened

new doors in dealing with complexity and uncertainty in AI explainability. Fuzzy systems can increase support and clarity in decision-making via managing uncertainty and explicating knowledge representations and inference. Garibaldi (2019) proposes a new framework for integrating fuzzy systems into AI explainability, which incorporates two systems: (1) technical fuzzy systems for representing and reasoning with ambiguity; and (2) fuzziness, for acknowledging and incorporating imperfection in AI decision-making and its explainability. However, research on the complementarity of AI explainability and fuzzy systems is still in its infancy, and more needs to be done to integrate fuzzy logic into explainability algorithms, to evaluate the effectiveness of fuzzy systems in dealing with uncertainty in the explainability domain and to examine the usefulness of outcomes of fuzzy AI explanations.

Finally, research is encouraged into how behavioral, societal and political factors affect AI explanation requirements and consequences. The first research stream is about a massive gap in theorizing the explainability and transparency of AI and particularly theorizing stakeholders' expectations of explanations in interactions with context-specific AI systems. Research to date has underutilized theories from social science, IS and business in conceptualizing the need for and the execution of explainability in AI systems, and thus more work is needed to build on existing theories and/or develop new theories about explainability effects. The second stream highlights that, apart from a few Chinese articles, all the papers selected in this review came from Western cultures, meaning less-developed countries are under-represented in AI knowledge. This is an important focus for future research; scholars like Felzmann *et al.* (2019) argue that human–computer studies of AI explainability need to be conducted in all regions because the contextual factors of countries and regions can affect the transferability of recommendations about transparency and explainability of AI systems.

6. Implications for theory and practice

6.1 Theoretical implications

This study employed contingency theory in a review of AI systems and expectations for their explainability. In line with the theory's original predictions (Feidler, 1964; Csaszar and Ostler, 2020), we found no universal or best way to implement AI explainability practices because the designs of AI systems need to fit the environments in which they are operating. The paper extends the theory's suggestions that the needs and goals of an organization in developing and using subsystems (i.e. AI systems) are better satisfied when its management style fits both the task (i.e. AI explanation) and the nature of the work (i.e. opposing effects) within the environment (i.e. the context in which the AI system operates). Our findings also respond to Tosi and Slocum's (1984) call for a balance between profitability, satisfaction and social responsibility, by presenting a framework that foregrounds the conflicting effects of AI explainability.

Comparing the opposing effects of AI explainability with contingency theory highlights the need for finding a mechanism for managing the trade-off between these effects in future AI developments and use. Until now, most practitioners and scholars have either overpraised high-performance AI systems and neglected their low visibility or focused solely on making these systems visible to humans and failed to appreciate their consequences properly. Hence, we searched the extant literature and revealed contradictory effects of explainability, which are defined as interrelated explainability elements that often co-exist simultaneously and persist over time when a black-box AI system is in use. We applied a contingency lens to the AI explainability domain, shedding light on the tensions between these opposing effects that span the explainability phenomena, analyzing its impacts on stakeholders and theoretical perspectives.

Furthermore, we presented a framework of six perspectives for managing the opposing effects of explainability by putting context at the center. These perspectives start with the

question of whether explainability is needed for a particular application/context. If so, they present a context-aware approach to managing unwanted effects of explainability. Thus, the framework offers the basis for the advancement of theorizing explainability of AI by providing a common understanding. At its core, it assumes that tensions are integral to AI systems and that successful, safe and compliant use of AI depends on attending to opposing yet intertwined demands simultaneously.

6.2 Practical implications

AI has changed dramatically over recent years. While AI mainly was used as a tool for automation or running complex calculations and organizational operations in its early days, it has become a productivity, hedonic and transformative tool for individual users and organizations. However, the explainability and transparency considerations of AI systems are not keeping pace with the technology's deployment rate. Fortunately, our findings reveal that the recent rise in AI ethical frameworks and governments' compliance expectations have redirected the focus from outcomes to tackling AI explainability, transparency and accountability issues. This highlights the pressing need for managers to develop AI explainability practices and, more importantly, for dealing with its unwanted and conflicting effects.

This study offers important practical implications for designing AI systems that are efficient and accurate while appreciative of compliance and contingent visibility requirements. AI is rapidly becoming a valuable tool supporting organizations' drive for growth and profitability, so vigilant management is needed to avoid unintentional or intentional damage to brand reputation and—more importantly—to users, other stakeholders and society as a whole. Thus, AI owners and developers are urged to look beyond profit, prediction accuracy and system speed; often, these numbers serve as hard evidence of AI success or failure, but they should not be regarded as the only indicators. Equally or even more important factors are making sound systems that lead to the creation of accurate algorithms and reliable recommendations. Still, they are harmless to humans, trustworthy, compliant with regulations and ethically sound.

Moreover, as governments, non-profit organizations and major technology firms develop ethical frameworks in response to demand for safe and transparent systems, AI practitioners and developers need to find a balance between ensuring systems' performance and explainability and accountability. This is important for long-term sustainability and can reduce the implications of making systems compliant. Lastly, standards organizations and governments and non-profit agencies need to accelerate the creation of metrics and standards for quantifying explainability expectations and outcomes. There is currently significant demand from society and the market for subjective and integrated standards and measurement tools for general and context-specific purposes that can be used to examine the explainability of AI and guide compliance policymaking while considering the confidentiality, performance and inherent complexity of these systems.

7. Conclusion and limitations

We reviewed and synthesized research in transparency, interpretability and explainability of AI systems and unpacked the tensions between five opposing effects of the explainability of these systems. Our results attest to the importance of taking a balanced standpoint in the quest for explainable AI systems so that benefits and the pressing need for AI visibility do not damage legitimate and inherent expectations of these systems. Clearly, some stakeholders of some AI systems need and experience the beneficial effects of explainable AI systems. However, benefits for some other stakeholders might be not so clear; explainability can backfire since it may prioritize the transparency of the system over understanding the problem space, create false binaries, or lead to harmful outcomes.

In this research, we recognized and identified five perspectives that can help manage the tension between the opposing effects of explainability in AI. We contend that context-aware consideration and execution of these perspectives can allow organizations and societies to take advantage of AI systems while controlling harmful or unwanted consequences. We argue that the starting point is a pragmatic definition of explainability, followed by consideration of the context in which explanation is needed. Next, to help organizations and decision-makers more efficiently conceptualize and consider the context in human–AI interactions, we introduced the notion of cohabitation of human agency and AI agency. We discussed its implications for how the responsibility for the outputs of advanced AI systems can be shared between both parties. We then discussed the lack and importance of a systematic and comprehensive set of measures for interpreting rights and wrongs in explanations in AI and how local or global explanation principles and a framework can help with such interpretations. Finally, we presented emerging technical solutions for improving the efficiency of explainability of AI that are receiving attention from scholars and practitioners.

Explainability of AI is a complex problem. Hence, specific explainability practices, expectations and techniques may not (and perhaps should not) be expected to be consistently generalizable to all systems and all applications. We argue that explainability is context-specific, and compromises between and within the 5C effects need to be made depending on the context in which AI operates. A wide variety of contexts exist, including geographical, cultural, organizational and human characteristics (Davison and Martinsons, 2016). While in a commercial context, in a particular geographical region, explainability of the AI system may be sacrificed in favor of performance and confidentiality, confidence in the AI system may be a top priority for a non-profit health entity in another geographical region. We concur with Davison and Martinsons's (2016) view that human and institutional differences matter in generalizing findings and imposing boundary conditions on AI explainability practices and expectations. Thus, we urge future researchers to study contextual factors that influence the encouragement or discouragement of explainability practices.

Because these findings are based on the literature, the limitations of the reviewed papers also apply to this research. For example, some papers were not found in our search and thus were not included in this study. Further to this, given the current pace of research in AI, there may be relevant studies that have been published after our search date and hence did not inform this review.

References

Abedin, B., Milne, D. and Erfani, E. (2020), "Attraction, selection, and attrition in online health communities: initial conversations and their association with subsequent activity levels", *International Journal of Medical Informatics*, Vol. 141, p. 104216.

Adadi, A. and Berrada, M. (2018), "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)", *IEEE Access*, Vol. 6, pp. 52138-52160.

Ågerfalk, P.J. (2020), "Artificial intelligence as digital agency", *European Journal of Information Systems*, Vol. 29 No. 1, pp. 1-8.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D. and Benjamins, R. (2020), "Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82-115.

Australian Government (2019), "AI ethics framework", available at: https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework (accessed 20 June 2020).

Bellotti, V. and Edwards, K. (2001), "Intelligibility and accountability: human considerations in context-aware systems", *Human–Computer Interaction*, Vol. 16 Nos 2-4, pp. 193-212.

Berente, N., Gu, B., Recker, J. and Santhanam, R. (2019), "Managing AI", *Management of Information Systems Quarterly*, available at: https://www.misq.org.misq.pdf.CurrentCalls (accessed 01 February 2021).

Berkelaar, B.L. (2014), "Cybervetting, online information, and personnel selection: new transparency expectations and the emergence of a digital social contract", *Management Communication Quarterly*, Vol. 28 No. 4, pp. 479-506.

Bertino, E., Kundu, A. and Sura, Z. (2019), "Data transparency with blockchain and AI ethics", *Journal of Data and Information Quality*, Vol. 11 No. 4, pp. 1-8.

Beydoun, G., Abedin, B., Merigó, J.M. and Vera, M. (2019), "Twenty years of information systems frontiers", *Information Systems Frontiers*, Vol. 21 No. 2, pp. 485-494.

Biran, O. and Cotton, C. (2017), "Explanation and justification in machine learning: a survey", *International Joint Conference on Artificial Intelligence*, Melbourne, Australia.

Chen, Q., Feng, Y., Liu, L. and Tian, X. (2019), "Understanding consumers' reactance of online personalized advertising: a new scheme of rational choice from a perspective of negative effects", *International Journal of Information Management*, Vol. 44, pp. 53-64.

Coeckelbergh, M. (2009), "Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents", *AI and Society*, Vol. 24 No. 2, pp. 181-189.

Coeckelbergh, M. (2020), "Artificial intelligence, responsibility attribution, and a relational justification of explainability", *Science and Engineering Ethics*, Vol. 26, pp. 2051-2068.

Csaszar, F.A. and Ostler, J. (2020), "A contingency theory of representational complexity in organizations", *Organization Science*, Vol. 31 No. 5, pp. 1053-1312.

Davison, R.M. and Martinsons, M.G. (2016), "Context is king! Considering particularism in research design and reporting", *Journal of Information Technology*, Vol. 31 No. 3, pp. 241-249.

De Graaf, M.M. and Malle, B.F. (2017), "How people explain action (and autonomous intelligent systems should too)", *Proceedings of The Association for the Advancement of Artificial Intelligence*, Virginia, USA, pp. 19-26.

De Wever, B., Schellens, T., Valcke, M. and Van Keer, H. (2006), "Content analysis schemes to analyze transcripts of online asynchronous discussion groups: a review", *Computers and Cducation*, Vol. 46 No. 1, pp. 6-28.

Diez-Olivan, A., Del Ser, J., Galar, D. and Sierra, B. (2019), "Data fusion and machine learning for industrial prognosis: trends and perspectives towards Industry 4.0", *Information Fusion*, Vol. 50, pp. 92-111.

Doshi-Velez, F. and Kim, B. (2017), *Towards a Rigorous Science of Interpretable Machine Learning*, arXiv preprint arXiv:1702.08608.

Dubé, L. and Robey, D. (2009), "Surviving the paradoxes of virtual teamwork", *Information Systems Journal*, Vol. 19 No. 1, pp. 3-30.

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., lavarasank, P.V., Janssenl, M., Jonesm, P., Kark, A.K., Kizginb, H., Kronemannm, B., Lal, B. and Williams, M.D. (2019), "Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy", *International Journal of Information Management*, Vol. 57, p. 101994.

D'Acquisto, G. (2020), "On conflicts between ethical and logical principles in artificial intelligence", *AI and Society*, Vol. 35, pp. 895-900.

Erfani, S.S. and Abedin, B. (2018), "Impacts of the use of social network sites on users' psychological well-being: a systematic review", *Journal of the Association for Information Science and Technology*, Vol. 69 No. 7, pp. 900-912.

European Commission (2018), "Communication from the commission to the European parliament, the European council, the council, the European economic and social committee and the committee

of the regions, artificial intelligence for Europe", available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN (accessed 01 February 2021).

European Commission (2019), "Ethics guidelines for trustworthy AI", available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 01 February 2020).

European Commission (2020), "COM(2010) 2020: Europe 2020—a strategy for Smart, Sustainable and inclusive growth", available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Aem0028 (accessed 01 February 2021).

Feidler, F. (1964), *Advances in Experimental Social Psychology, Contingency Model of Leadership Effectiveness*, Academic Press, New York, NY.

Felzmann, H., Villaronga, E.F., Lutz, C. and Tamò-Larrieux, A. (2019), "Transparency you can trust: transparency requirements for artificial intelligence between legal Norms and contextual concerns", *Big Data and Society*, Vol. 6 No. 1, pp. 1-14.

Fernandez, A., Herrera, F., Cordon, O., del Jesus, M.J. and Marcelloni, F. (2019), "Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to?", *IEEE Computational Intelligence Magazine*, Vol. 14 No. 1, pp. 69-81.

Floridi, L. (2011), "Children of the fourth revolution", *Philosophy and Technology*, Vol. 24 No. 3, pp. 227-232.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V. and Rossi, F. (2018), "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations", *Minds and Machines*, Vol. 28 No. 4, pp. 689-707.

Galbraith, J.R. (1973), *Designing Complex Organizations*, Addison-Wesley Longman Publishing, MA.

Garibaldi, J.M. (2019), "The need for fuzzy AI", *IEEE/CAA Journal of Automatica Sinica*, Vol. 6 No. 3, pp. 610-622.

Gödel, K. (1931), "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme", *Monatshefte für Mathematik und Physik*, Vol. 38 No. 1, pp. 173-198.

Goodman, B. and Flaxman, S. (2017), "European Union regulations on algorithmic decision-making and a 'right to explanation'", *AI Magazine*, Vol. 38 No. 3, pp. 50-57.

Gregor, S. and Benbasat, I. (1999), "Explanations from intelligent systems: theoretical foundations and implications for practice", *MIS Quarterly*, Vol. 23 No. 4, pp. 497-530.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018), "A survey of methods for explaining black box models", *ACM Computing Surveys*, Vol. 51 No. 5, pp. 1-42.

Gunning, D. and Aha, D.W. (2019), "DARPA's explainable artificial intelligence program", *AI Magazine*, Vol. 40 No. 2, pp. 44-58.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G.-Z. (2019), "XAI—explainable artificial intelligence", *Science Robotics*, Vol. 4 No. 37, p. 7120.

Guzman, A.L. and Lewis, S.C. (2020), "Artificial intelligence and communication: a Human–Machine communication research agenda", *New Media and Society*, Vol. 22 No. 1, pp. 70-86.

Hacker, P., Krestel, R., Grundmann, S. and Naumann, F. (2020), "Explainable AI under contract and tort law: legal incentives and technical challenges", *Artificial Intelligence and Law*, Vol. 28, pp. 415-439.

Harper, R.H. (2019), "The role of HCI in the age of AI", *International Journal of Human–Computer Interaction*, Vol. 35 No. 15, pp. 1331-1344.

Hoffman, R.R. and Klein, G. (2017), "Explaining explanation, part 1: theoretical foundations", *IEEE Intelligent Systems*, Vol. 32 No. 3, pp. 68-73.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H. (2019), "Causability and explainability of artificial intelligence in medicine", *Data Mining and Knowledge Discovery*, Vol. 9 No. 4, e1312.

Kaplan, A. and Haenlein, M. (2019), "Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence", *Business Horizons*, Vol. 62 No. 1, pp. 15-25.

Kitchenham, B. (2004), "Procedures for performing systematic reviews", *Keele*, Vol. 33 No. 2004, pp. 1-26.

Kitchenham, B. (2007), *Guidelines for Performing Systematic Literature Reviews in Software Engineering*, Technical report, Ver. 2.3 EBSE Technical Report, *EBSE*: Keele University and University of Durham, available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.471&rep=rep1&type=pdf (accessed 01 February 2020).

Krippendorff, K. (1989), "Content analysis", *International Encyclopedia of Communication*, Vol. 1, pp. 403-407.

Kroll, J.A. (2018), "Data science data governance [AI ethics]", *IEEE Security and Privacy*, Vol. 16 No. 6, pp. 61-70.

Kshetri, N. (2019), "Complementary and synergistic properties of blockchain and artificial intelligence", *IT Professional*, Vol. 21 No. 6, pp. 60-65.

Lawless, W.F., Mittu, R., Sofge, D. and Hiatt, L. (2019), "Artificial intelligence, autonomy, and human-machine teams: interdependence, context, and explainable AI", *AI Magazine*, Vol. 40 No. 3, pp. 5-13.

Lecue, F. (2019), "On the role of knowledge graphs in explainable AI", *Semantic Web*, Vol. 11 No. 1, pp. 41-51.

Lewis, M.W. (2000), "Exploring paradox: toward a more comprehensive guide", *Academy of Management Review*, Vol. 25 No. 4, pp. 760-776.

MacDonell, S., Shepperd, M., Kitchenham, B. and Mendes, E. (2010), "How reliable are systematic reviews in empirical software engineering?", *IEEE Transactions on Software Engineering*, Vol. 36 No. 5, pp. 676-687.

Majchrzak, A., Faraj, S., Kane, G.C. and Azad, B. (2013), "The contradictory influence of social media affordances on online communal knowledge sharing", *Journal of Computer-Mediated Communication*, Vol. 19 No. 1, pp. 38-55.

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

Miller, T. (2019), "Explanation in artificial intelligence: insights from the social sciences", *Artificial Intelligence*, Vol. 267, pp. 1-38.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016), "The ethics of algorithms: mapping the debate", *Big Data and Society*, Vol. 3 No. 2, pp. 1-21.

Mohamed, S., Png, M.T. and Isaac, W. (2020), "Decolonial ai: decolonial theory as sociotechnical foresight in artificial intelligence", *Philosophy and Technology*, Vol. 33, pp. 659-684.

Mohseni, S. and Ragan, E.D. (2018), *A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning*, arXiv preprint arXiv:1801.05075.

Mohseni, S., Zarei, N. and Ragan, E.D. (2018), *A Survey of Evaluation Methods and Measures for Interpretable Machine Learning*, arXiv preprint arXiv:1811.11839.

Nassar, M., Salah, K., Rehman, M.H. and Svetinovic, D. (2020), "Blockchain for explainable and trustworthy artificial intelligence", *Data Mining and Knowledge Discovery*, Vol. 10 No. 1, e1340.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E. and Krasanakis, E. (2020), "Bias in data-driven artificial intelligence systems—an introductory survey", *Data Mining and Knowledge Discovery*, Vol. 10 No. 3, e1356.

Odlyzko, A. (2019), "Cybersecurity is Not very important", *Ubiquity*, Vol. 2019, June, pp. 1-23.

Páez, A. (2019), "The pragmatic turn in explainable artificial intelligence (XAI)", *Minds and Machines*, Vol. 29 No. 3, pp. 441-459.

Pieters, W. (2011), "Explanation and trust: what to tell the user in security and AI?", *Ethics and Information Technology*, Vol. 13 No. 1, pp. 53-64.

Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W. and Wallach, H. (2018), *Manipulating and Measuring Model Interpretability*, arXiv preprint arXiv:1802.07810.

Preece, A. (2018), "Asking 'Why' in AI: explainability of intelligent systems–perspectives and challenges", *Intelligent Systems in Accounting, Finance and Management*, Vol. 25 No. 2, pp. 63-72.

Priharsari, D., Abedin, B. and Mastio, E. (2020), "Value co-creation in firm sponsored online communities", *Internet Research*, Vol. 30 No. 3, pp. 763-788.

Rai, A. (2020), "Explainable AI: from black box to glass box", *Journal of the Academy of Marketing Science*, Vol. 48 No. 1, pp. 137-141.

Reinking, J. (2012), "Contingency theory in information systems research", *Information Systems Theory*, Vol. 28, pp. 247-263.

Robbins, S. (2019a), "AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines", *AI and Society*, Vol. 35, pp. 391-400.

Robbins, S. (2019b), "A misdirected principle with a catch: explicability for AI", *Minds and Machines*, Vol. 29 No. 4, pp. 495-514.

Schad, J., Lewis, M.W., Raisch, S. and Smith, W.K. (2016), "Paradox research in management science: looking back to move forward", *The Academy of Management Annals*, Vol. 10 No. 1, pp. 5-64.

Shao, Z., Feng, Y. and Hu, Q. (2016), "Effectiveness of top management support in enterprise systems success: a contingency perspective of fit between leadership style and system life-cycle", *European Journal of Information Systems*, Vol. 25 No. 2, pp. 131-153.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Lanctot, M. (2016), "Mastering the game of Go with deep neural networks and tree search", *Nature*, Vol. 529 No. 7587, pp. 484-489.

Smith, W.K. and Lewis, M.W. (2011), "Toward a theory of paradox: a dynamic equilibrium model of organizing", *Academy of Management Review*, Vol. 36 No. 2, pp. 381-403.

Stahl, B.C. and Wright, D. (2018), "Ethics and privacy in AI and big data: implementing responsible research and innovation", *IEEE Security and Privacy*, Vol. 16 No. 3, pp. 26-33.

Timmers, P. (2019), "Ethics of AI and cybersecurity when sovereignty is at stake", *Minds and Machines*, Vol. 29 No. 4, pp. 635-645.

Tóth, A.K. (2019), "Algorithmic copyright enforcement and AI: issues and potential solutions through the lens of text and data mining", *Masaryk University Journal of Law and Technology*, Vol. 13 No. 2, pp. 361-388.

Tosi, H.L. Jr and Slocum, J.W. Jr (1984), "Contingency theory: some suggested directions", *Journal of Management*, Vol. 10 No. 1, pp. 9-26.

Turing, A.M. (2009), "Computing machinery and intelligence", *Parsing the Turing Test*, pp. 23-65.

Waltl, B. and Vogl, R. (2018), "Explainable artificial intelligence- the new Frontier in legal informatics", *Jusletter IT*, Vol. 4, pp. 1-10.

Wang, D., Yang, Q., Abdul, A. and Lim, B. Y. (2019), "Designing theory-driven user-centric explainable AI", *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-15.

Weill, P. and Olson, M.H. (1989), "An assessment of the contingency theory of management information systems", *Journal of Management Information Systems*, Vol. 6 No. 1, pp. 59-86.

Wolfswinkel, J.F., Furtmueller, E. and Wilderom, C.P. (2013), "Using grounded theory as a method for rigorously reviewing literature", *European Journal of Information Systems*, Vol. 22 No. 1, pp. 45-55.

Zadeh, L.A. (1965), "Fuzzy sets", *Information and Control*, Vol. 8 No. 3, pp. 338-353.

Zhang, L. (2019), "China: AI governance principles released", available at: https://www.loc.gov/law/foreign-news/article/china-ai-governance-principles-released/ (accessed 09 May 2021).

Appendix

Authors/year	Title	Journal
Adadi and Berrada (2018)	Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)	IEEE access
Arrieta *et al.* (2020)	Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI	Information fusion
Berkelaar (2014)	Cybervetting, online information, and personnel selection: New transparency expectations and the emergence of a digital social contract	Management communication quarterly
Bertino *et al.* (2019)	Data transparency with blockchain and AI ethics	Journal of data and information quality
Coeckelbergh (2009)	Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents	AI and society
Coeckelbergh (2020)	Artificial intelligence, responsibility attribution, and a relational justification of explainability	Science and engineering ethics
D'Acquisto (2020)	On conflicts between ethical and logical principles in artificial intelligence	AI and society
Doshi-Velez and Kim (2017)	Towards a rigorous science of interpretable machine learning	arXiv preprint arXiv:1702.08608
Felzmann *et al.* (2019)	Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns	Big data and society
Fernandez *et al.* (2019)	Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?	IEEE computational intelligence magazine
Floridi (2011)	Children of the fourth revolution	Philosophy and technology
Floridi *et al.* (2018)	AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations	Minds and machines
Garibaldi (2019)	The need for fuzzy AI	IEEE/CAA journal of automatica sinica
Gunning and Aha (2019)	DARPA's explainable artificial intelligence program	AI magazine
Gunning *et al.* (2019)	XAI—explainable artificial intelligence	Science robotics
Hacker *et al.* (2020)	Explainable AI under contract and tort law: legal incentives and technical challenges	Artificial intelligence and law
Harper (2019)	The role of HCI in the age of AI	International journal of human–computer interactions
Holzinger *et al.* (2019)	Causability and explainability of artificial intelligence in medicine	Data mining and knowledge discovery
Kaplan and Haenlein (2019)	Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence	Business horizons
Kroll (2018)	Data science data governance [AI ethics]	IEEE security and privacy
Kshetri (2019)	Complementary and synergistic properties of blockchain and artificial intelligence	IT professional
Lawless *et al.* (2019)	Artificial intelligence, autonomy, and human-machine teams: Interdependence, context, and explainable AI	AI magazine
Lecue (2019)	On the role of knowledge graphs in explainable AI	Semantic web

Table A1.
Papers selected for the review

(*continued*)

Authors/year	Title	Journal
Miller (2019)	Explanation in artificial intelligence: Insights from the social sciences	Artificial intelligence
Mittelstadt *et al.* (2016)	The ethics of algorithms: Mapping the debate	Big data and society
Mohseni and Ragan (2018)	A human-grounded evaluation benchmark for local explanations of machine learning	arXiv preprint arXiv:1801.05075
Nassar *et al.* (2020)	Blockchain for explainable and trustworthy artificial intelligence	Data mining and knowledge discovery
Ntoutsi *et al.* (2020)	Bias in data-driven artificial intelligence systems—An introductory survey	Data mining and knowledge discovery
Páez (2019)	The pragmatic turn in explainable artificial intelligence (XAI)	Minds and machines
Pieters (2011)	Explanation and trust: What to tell the user in security and AI?	Ethics and information technology
Poursabzi-Sangdeh *et al.* (2018)	Manipulating and measuring model interpretability	arXiv preprint arXiv:1802.07810
Preece (2018)	Asking "why" in AI: Explainability of intelligent systems–perspectives and challenges	Intelligent systems in accounting, finance and management
Rai (2020)	Explainable AI: from black box to glass box	Journal of academy of marketing science
Robbins (2019a)	AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines	AI and society
Robbins (2019b)	A misdirected principle with a catch: Explicability for AI	Minds and machines
Stahl and Wright (2018)	Ethics and privacy in AI and big data: Implementing responsible research and innovation	IEEE security and privacy
Timmers (2019)	Ethics of AI and cybersecurity when sovereignty is at stake	Minds and machines
Tóth (2019)	Algorithmic copyright enforcement and AI: Issues and potential solutions through the lens of text and data mining	Masaryk university journal of law and technology
Waltl and Vogl (2018)	Explainable artificial intelligence- the new frontier in legal informatics	Jusletter IT

Table A1.

Corresponding author
Babak Abedin can be contacted at: babak.abedin@mq.edu.au