**Trust or Mistrust in Algorithmic Grading? An Embedded Agency Perspective**

Stephen Jackson

Ontario Tech University, Oshawa, Canada

[stephen.jackson@uoit.ca](mailto:stephen.jackson@uoit.ca)

Niki Panteli

Royal Holloway, Egham, Surrey, TW20 0EX, United Kingdom & Norwegian University of Science and Technology, Trondheim, Norway

[niki.panteli@rhul.ac.uk](mailto:niki.panteli@rhul.ac.uk)

Keywords: Trust, Mistrust, Algorithmic Grading, Multi-Level Analysis, Embedded Agency

**Abstract**

Artificial Intelligence (AI) has the potential to significantly impact the educational sector. One application of AI that has increasingly been applied is algorithmic grading. It is within this context that our study takes a focus on trust. While the concept of trust continues to grow in importance among AI researchers and practitioners, an investigation of trust/mistrust in algorithmic grading across multiple levels of analysis has so far been under-researched. In this paper, we argue the need for a model that encompasses the multi-layered nature of trust/mistrust in AI. Drawing on an embedded agency perspective, a model is devised that examines top-down and bottom-up forces that can influence trust/mistrust in algorithmic grading. We illustrate how the model can be applied by drawing on the case of the International Baccalaureate (IB) program in 2020, whereby an algorithm was used to determine student grades. This paper contributes to the AI-trust literature by providing a fresh theoretical lens based on institutional theory to investigate the dynamic and multi-faceted nature of trust/mistrust in algorithmic grading—an area that has seldom been explored, both theoretically and empirically. The study raises important implications for algorithmic design and awareness. Algorithms need to be designed in a transparent, fair, and ultimately a trustworthy manner. While an algorithm typically operates like a black box, whereby the underlying mechanisms are not apparent to those impacted by it, the purpose and an understanding of how the algorithm works should be communicated upfront and in a timely manner.

**1. Introduction**

Organizations and societies worldwide are increasingly harnessing the potential of AI, as witnessed by the growth in AI technologies, tools, and applications (Alter, 2021; Hradecky et al., 2022; Wang, Teo, & Janssen, 2021). AI, put broadly, refers to intelligent machines that can behave and react similarly to the cognitive functions typically performed by humans. Some of these functions may include learning, reasoning, making inferences, judgments, predictions, complex decision-making, interaction, and communication (Davenport & Ronanki, 2018; Rai, Constantinides, & Sarker, 2019). AI technologies and applications have the potential to impact a broad range of business areas, including automation, robotics, machine learning, and emotion detection, to name a few (Anderson, Rainie, & Luchsinger, 2018; Dwivedi et al., 2021), and have been adopted by many different industries, including fintech, healthcare, transport, and information security. One sector that has not been immune to advancements in AI is the educational sector (Hwang et al., 2020). While, in comparison to other industries, the educational sector has been slow to embrace AI-based technologies and applications, it is anticipated that this will change. Vincent-Lancrin and van der Vlies (2020, p.6) note “while most innovation in the past decade related to an increased use of computers and the internet in the classroom, the next wave will be based on AI, or on combinations of AI and other technologies.”

For some, there are great expectations that AI will play a crucial role in transforming educational practices (Kaur, Tandon, & Matharou, 2020). From a student perspective, AI can provide personalization through smart content that tailors the learning experience to the needs of each student at a time of need (Hwang et al., 2020; Karandish, 2021; Plitnichenko, 2020). For example, the use of AI virtual assistants or chatbots can offer students 24/7 learning access and assist with queries within a real-time environment. From an organizational standpoint, AI can assist administrators in the mining of large datasets to identify trends and make performance predictions. This can help with business process optimization, improve business forecasting, as well as cut expenditure in the long run (Kolbjørnsrud, Amico, & Thomas, 2016). From the perspective of the teacher, AI can assist with the preparation of learning content, free up time to concentrate on other value-added activities, and support with grading (Goel & Joyner, 2017; Khare, Stewart, & Khare, 2018).

While the use of AI has the potential to bring educational benefits (Luckin et al., 2016), one application of educational AI that is gaining increased media attention is algorithmic grading (Jackson & Panteli, 2021). This attention has been spurred on by what can be referred to as the dark side of AI, where initially the intended purpose is to produce anticipated benefits, yet, in many cases, has been met with undesirable outcomes. Concerns around algorithmic bias and fairness, lack of transparency in terms of how AI functions and makes decisions, issues relating to privacy, as well as the propensity for students to cheat or find loopholes to manipulate the system, have been acknowledged (Bogina et al., 2021; Larsson & Heintz, 2020; Smith, 2020). Parsons (2020), for instance, reported how students were able to outwit the AI algorithm used to assign assessment grades. Realizing that the algorithm was merely scanning a series of keywords based on the topic of the assessment, students were able to game the system by inserting keywords, rather than providing a detailed answer.

An important issue that lies at the very heart of computerized applications including AI, affecting how they are successfully used and deployed, is trust. This is evidenced by the growth in articles focusing on trust in AI (e.g. Glikson & Woolley, 2020; Hengstler, Enkel, & Duelli, 2016; Ologeanu-Taddei & Vitari, 2020; Siau & Wang, 2018; Toreini et al., 2020). Regardless of the importance of trust in the adoption and use of AI, our understanding of trust or the lack of it, thus mistrust, in relation to the study of algorithmic grading, has been limited. For this study, algorithmic grading is defined as a set of rules or procedures undertaken by computerized systems that can be used to perform calculations or problem solving to assist with student grade prediction.

Furthermore, while research on trust and AI is still evolving, a closer examination of the literature revealed that there is a tendency for researchers to analyze trust at a single level, particularly the individual users, while the literature on trust incorporating multiple levels of analysis remains limited (Jackson & Panteli, 2020). There have been increased calls to understand further, both empirically and theoretically, the top-down and bottom-up forces that can influence trust processes, taking into consideration multiple levels, cross-level effects, and trust/mistrust dynamics (Lumineau & Schilke, 2018; Mollering, Gillespie, & Lewicki, 2021). Mollering et al., (2021, p.363), note that “a rich area for future research is the examination of whether and how top-down and bottom-up processes may operate in a particular context and around a specific form or dynamic of trust.” It is our position that with the increasing popularity of AI for educational purposes, these issues require further attention. Following from this, the aim of the study has been to understand the multi-layered nature of trust/mistrust in algorithmic grading. Therefore, the broad research question that drives the study is: ‘How do top-down and bottom-up forces across multiple levels influence trust/mistrust associated with algorithmic grading?’

In doing so, we draw on the embedded agency perspective (Lumineau & Schilke, 2018). Central to embedded agency is an examination of the reciprocal relationship between top-down and bottom-up forces at play, which can influence initial trust/mistrust formation and its continuous development. The paper argues that by considering multiple layers, and the activity that occurs across levels, it can provide rich insights into the AI phenomenon under investigation. The embedded agency approach is illustrated by drawing on the case of IB and its use of algorithmic grading in 2020. More specifically, based on the constructionist research paradigm, we examine various users’ responses who were, directly or indirectly, impacted by the grading algorithm from across the globe to illustrate the workings of the approach put forward. A key finding of this study is that although top-down forces, i.e. institutional managers, can shape individuals’ trust perceptions, and in the case of IB, promote mistrust in algorithmic grading among lower levels, individuals, e.g. students, parents, teachers, through collective action, can lead to revised trust processes at the organizational level.

In summary, the article makes several important contributions to the information systems (IS) literature. Firstly, although the concept of trust has been studied in the context of AI, there is an increased need to analyze trust in the case of algorithmic grading (Jackson & Panteli, 2021). Given the likely continued and increased adoption of AI in education, as well as the dearth of studies focusing exclusively on algorithmic grading, we feel that this area is worthy of investigation. Second, the paper provides fresh theoretical insights by adopting an embedded agency perspective based on institutional theory to the study of algorithmic grading—a theoretical perspective that has received less consideration in the AI-trust literature. Thirdly, moving beyond the treatment of trust as a singular phenomenon of interest, in this study, we position our focus on trust as a multi-faceted, dynamic, and emergent phenomenon that is constructed and reconstructed by individuals over a period of time, depending on the situational context in which the AI-application is embedded. A consideration of trust in relation to issues pertaining to layers, dynamism, context, and time, which are central themes in embedded agency, may lead to fresh insights into trust/mistrust formation practices and open up new lines of theoretical and methodological inquiry.

The results of this study may benefit those responsible for developing and deploying AI in a number of ways. As algorithms continue to be used in the educational sector, it is important that they are designed and developed in a way that is fair and transparent to facilitate stakeholder trust. Institutional managers should be mindful that their inactions, e.g. poor decision-making processes, lack of effective communication, and taking a highly technocratic approach, can foster stakeholder discontentment and mistrust in AI. Attention and effort should be given to understanding the expectations, intentions, and attitudes of those impacted by AI throughout all stages of its lifecycle. These elements should be considered within the context in which the algorithm exists.

This paper is organized as follows. In the next section (Section 2), the concepts of AI and trust are introduced and explored more fully, drawing particularly on the organizational and IS literature. Next, in Section 3, the theoretical framework (an embedded agency perspective) is outlined, including its application to the study of AI and trust. Next, a discussion of the research methods and data analysis techniques used for this study are outlined (Section 4). The findings based on the case of the IB algorithm to determine student grades are then provided in Section 5. This is followed by a discussion of the findings, implications, limitations, and suggestions for future research of the study in Section 6. Section 7 presents the key conclusions of the study.

**2. AI and trust**

Educational institutions, like many organizations, are increasingly contemplating or already harnessing the potential of AI (Bonderud, 2019). According to Emergen Research (2022), the international market for AI in education is anticipated to reach US$17.83 billion by 2027. There have been great expectations that AI will play a crucial role in transforming the educational sector. Proponents, for instance, have acknowledged the benefits of AI for enhancing the overall student experience, optimizing business processes, and driving through improvements in efficiencies (Kavitha et al., 2018; Nalbant, 2021).

Although on the surface, the adoption of AI in general may signal increased uptake, the potential impact that AI can have on learning and teaching is still not clear (Bates et al., 2020; Popenici & Kerr, 2017; Schiff, 2021; Zawacki-Richter et al., 2019). As noted by Bates et al. (2020, p.4), “AI is in widespread use in some areas of society. In its direct impact on teaching and learning though, much has been promised, but as yet, little has been achieved.” This raises an interesting question – why are many educational institutions not reaping the transformational potential with AI as first envisioned? Although technical issues in the development and implementation of AI technologies continue to act as a barrier to successful adoption, many of the current challenges faced when introducing AI are social and behavioral in nature. Some of these challenges relate to privacy, security, ethics, human rights, and expectations management (Chatterjee & Bhattacharjee, 2020; Zawacki-Richter et al., 2019).

One important factor in the use and deployment of AI is trust (Daugherty, Wilson, & Chowdhury, 2019; Gille, Jobin, & Ienca, 2020; Glikson & Woolley, 2020). Numerous scholars have agreed that trust is highly beneficial for AI adoption, satisfaction, or engagement (Chandra, Shirish, & Srivastava, 2021; Gille et al., 2020; Sethumadhavan, 2019; Stanton & Jensen, 2021). Siau and Wang (2018, p.47), for instance, acknowledged that trust is an essential element in encouraging uptake and use of AI. The authors stressed that in the case of AI, “both initial trust formation and continuous trust development deserve special attention.” In the study of AI-trust, one body of work, what we label as AI-centric, has focused on the technological features of AI (e.g. usability, accuracy, efficiency, robustness, performance) that foster trust (Bitkina et al., 2020; Papenmeier, Englebienne, & Seifert, 2019; Shin, 2021). Papenmeier et al. (2019), in studying the effects of AI accuracy and explainability on user trust, showed that accuracy was more important for establishing user trust than explainability. Liu (2021), in investigating human-AI interaction, found that system transparency, i.e. when the system outlined the details for its decisions, and the decisions of the AI were sound, enhanced trust, and reduced uncertainty.

Although the technical aspects of AI can affect trust (Sethumadhavan, 2019), other perspectives have focused on the human qualities and traits (e.g. personality types, cultural backgrounds, the abilities and competencies of users, user expectations, and experiences) that influence an individual’s, or perhaps more amply, the trustees’, disposition to trust AI. We refer to this as a human-centric position on AI-trust. The trustworthiness of an AI system, and the extent to which it is successfully deployed and adopted, depends on how it is ultimately perceived by the user (Hoff & Bashir, 2015). As noted by Stanton and Jensen (2021, p.1), “many current efforts are aimed to assess AI system trustworthiness through measurements of accuracy, reliability, and explainability, among other system characteristics. While these characteristics are necessary, determining that the AI system is trustworthy because it meets its system requirements won’t ensure widespread adoption of AI. It is the user, the human affected by the AI, who ultimately places their trust in the system.” Sethumadhaven (2019) acknowledged that in order to nurture the desired level of trust in AI, there is a need to consider a range of user characteristics, including age, culture, personality, gender, mood, self-confidence, and memory capacity. Bockle, Yeboah-Antwi, & Kouris (2021), in recognizing the lack of research focusing on trust, and particularly how AI-enabled interfaces are perceived by users, found a correlation between various personality type traits, e.g. extraversion, agreeableness, and open-mindedness, and the trust placed in interfaces.

What must also be considered is how the actions of different parties involved in AI development and deployment, can influence the trust process. Abbass (2019, p.167), in examining the issue of human-AI trust, based on the work of Mayer, Davis, and Schorman (1995), defined trust as the “willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.” A high degree of trust, as confident and optimistic anticipation in the actions and behavior of another stakeholder, is important for facilitating transparency.

Transparency plays an important role in nurturing user trust and integrity in AI. When users see an algorithm as being transparent, explainable, accountable, and fair, they are more inclined to perceive it as useful and trustworthy (Ryan, 2020; Shin, 2020). Thus, those responsible for deploying AI should be able to clearly justify the intentions for needing it, as well as provide affected parties with a basic understanding of its underlying functions or logic (Glikson & Woolley, 2020; Siau & Wang, 2018). The centrality of trust is further accentuated by its absence: mistrust makes AI success more difficult to attain (Kaplan, 2016). Ryan (2020) argued that AI can weaken the importance of interpersonal trust and removes the accountability of those responsible for developing and deploying it. Mistrust can result in users ignoring the recommendations suggested by AI, e.g. reverting to the use of human decision-making processes, as well as having an unconstructive impact on AI adoption (Cannon, 2019; Jackson & Panteli, 2020, 2021; Polonski, 2018).

Given that AI can perform tasks that mimic human intelligence and has the capacity for self-learning, the requirement to integrate the study of AI with issues relating to trust, as well as the need for fresh theoretical frameworks and empirical insights, has never been so important. However, with few notable exceptions (e.g. Glikson & Woolley, 2020; Hengstler, Enkel, & Duelli, 2016; Ologeanu-Taddei & Vitari, 2020; Siau & Wang, 2018), the study of trust in the context of AI is still in need of further investigation. While technology- and human-centric accounts have been useful in advancing our understanding of AI-trust, one area in need of further investigation is the dynamic nature of trust in AI. From this perspective, the relationship between trust and AI is active and complex, involving the entanglement of social, technological, and cognitive elements (Castelfranchi & Falcone, 2010; Ferrario, Loi, & Vigano, 2020). Asan, Bayrak, & Choudhury (2020) acknowledged that human-AI collaboration should be considered as a dynamic socio-technical system. Trust in AI, for example, can be influenced by social factors, but the social aspects can also be shaped by the AI.

Notwithstanding the importance of dynamic perspectives, research remains limited in this regard. Furthermore, the limited literature on trust and AI has focused exclusively on the individual user (Ologeanu-Taddei & Vitari, 2020), i.e. how an individual’s trust is affected by the functionality and reliability of AI technology. Accordingly, research on trust incorporating multiple levels of analysis remains incomplete (Fulmer & Dirks, 2018; Jackson & Panteli, 2020). Therefore, in order to gain further insight on trust/mistrust in relation to the study of algorithmic grading, research should be extended to include a multi-level perspective. This paper acknowledges that in order for researchers to understand how trust/mistrust influences algorithmic grading uptake and use, among other forms of AI, the focus of attention needs to shift from solely concentrating on just one level to also understanding the contextual, multi-faceted and dynamic nature of trust across multiple levels. A multi-level perspective of trust/mistrust in AI could more accurately explain human behavior (Glikson & Woolley, 2020). Our study builds on this AI-human dynamic perspective. A description of three different perspectives on AI-trust is captured in Table 1 below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AI-Centric** | **Human-Centric** | **AI-Human Dynamism** |
| Focus  Question of interest  Characteristics | Emphasis is on AI characteristics  What technology characteristics must AI possess in order to be considered as being trustworthy?  Predictable Deterministic  Stable | Emphasis is on human traits and qualities  What human qualities are desirable to nurture trust in AI?  Predictable  Deterministic  Stable | Emphasis is on the interplay between AI and human factors  How does the dynamic interplay between AI and human elements influence trust?  Unpredictable  Non-deterministic  Dynamic  Sensitive to time, space, and context |

**Table 1.** Perspectives on AI-trust

**3. An embedded agency perspective**

In this section, we outline a model based on an embedded agency perspective to examine trust/mistrust across multiple levels of analysis (Lumineau & Schilke, 2018). The model is formulated out of the concern that researchers tend to study one level, rather than accounting for multiple levels simultaneously (Erez & Gati, 2004; Klein & Kozlowski, 2000; Lumineau & Schilke, 2018). A key assumption is that considering the phenomena of interest from just one level can potentially, although not always, lead to a restrictive view. As trust is fundamentally about inter-related social relations (as outlined in Section 2), it is difficult not to consider trust as consisting of social relations influenced by other social relations across different levels and time.

*3.1. Embedded agency*

Embedded agency is an important part of institutional theory. Institutional theory, drawing on organizational and sociology literature, seeks to understand how structure (including routines, rules, norms, and beliefs) becomes established, and how structure can condition forms of social behavior (Scott, 2004; 2008). Much of the focus of earlier work using institutional theory sought to understand the specific institutional pressures, including normative, coercive, and mimetic pressures and mechanisms through which institutional isomorphism takes place. Isomorphism explains how organizations operating within a field embrace homogeneity, for instance, in relation to similar processes, rules, and systems (DiMaggio & Powell, 1983; Meyer & Rowan, 1977).

Moving away from the sole focus on the collective and established nature of structures, other researchers (Abdelnour, Hasselbladh, & Kallinikos, 2017; Seo & Creed, 2002) have sought to integrate notions of change and agency. Consequently, this has given rise to a dilemma that can be referred to as the “structure-agency debate” (Garud, Hardy, & Maguire, 2007), or the “paradox of embedded agency” (de Lange, 2019; Seo & Creed, 2002). The paradox of embedded agency relates to the “problem of accounting for structural change when the actions of agents are themselves conditioned by social structures that they wish to change” (Kholeif & Jack, 2019, p.62).

Rather than structure and agency being considered as separate, agency is fundamentally caught up and part of the very structure that actors themselves have formed or enacted. While agents (actors) are subject to pressures that structure their norms, beliefs, and routines, and appear initially to be fixed and institutionalized, actors also have the potential to exhibit agency—they have the capacity to reflect, challenge and adapt to new norms, and bring about change (Lawrence, Suddaby, & Leca, 2009; Seo & Creed, 2002; Wijen & Ansari, 2007; Zeitsma & Lawrence, 2010). Agents may join forces with other like-minded individuals who share similar beliefs and norms, and, through influence and pressure, can encourage existing norms to be challenged, changed, or re-established. Agency is not a rigid phenomenon; rather, it is best considered as temporal in nature, which is influenced by the context in which an individual finds oneself.

*3.2. Embedded agency and trust*

This section puts forward an embedded agency perspective on trust to examine top-down and bottom-up forces at play that influence trust/mistrust development. An embedded agency perspective on trust takes into consideration how top-down forces can influence trust/mistrust at lower levels, but also how lower-level activity can influence trust/mistrust processes at the higher level (Möllering, 2005). More specifically, organizational factors, guided by institutional norms, values, and practices, can influence trust processes at the group and individual levels, yet micro-level activity and alterations occurring at the bottom can also influence actions and trust processes at the top (organizational level). Thus, rather than focusing exclusively on one level, i.e. macro or micro levels, the focus is on the dynamic interaction when different levels come into contact and collide.

Trust is dynamic and evolving in nature, arising out of the interaction between top-down and bottom-up forces at play (Welter, 2012). In discussing the issue of trust, Lumineau and Schilke (2018, p.5) noted that “trust is inherently a multi-level phenomenon and, thus, that our understanding of trust development should embrace the reciprocal relationships between micro and macro perspectives.” Moreover, trust/mistrust formation and its continuous development are closely tied to its situational context. In order to comprehend trust/mistrust more fully, it needs to be considered within its broader historical, social, cultural, or community setting. We propose that the interactions between top-down and bottom-up forces are influenced by the situation within which they are in. A multi-level model, which largely considers trust/mistrust development across levels and at different stages of AI development and use is useful as it can be applied to understand the multi-faceted nature of trust/mistrust, particularly the dynamic interplay between macro and micro levels.

*3.2.1. Top-down forces*

Top-down forces at the organizational level, through formal and informal organizational structures, can influence lower-level activity. According to Lumineau and Schilke (2018, p.8), organizational structures are understood broadly as “both formal organizational design and informal organizational norms and procedures.” Thus, rather than structures being solely viewed as a system to classify hierarchy (e.g. job, role, and function) within an organization, it also consists of policies and norms that provide guidance in terms of decision-making processes and how decisions should be communicated with relevant stakeholders.

The norms, rules, practices, and policies of the organization can act as blueprints, offering cues and guidance in terms of how its members behave and think, as well as influence how individuals perceive trust, and what they deem as trustworthy behavior. Forces operating at the higher level can shape cognition and behaviors, as well as enable or constrain trust at lower levels. In order to enable trust, Whitener et al. (1998, p.514) note that “managers’ actions and behaviors provide the foundation for trust and that it is actually management’s responsibility to take the first step and initiate trusting relationships.”

As a way of nurturing trust, managers should endorse fairness, openness, integrity, honesty, and consistency (Axelrod, 1984; Roberts & O’Reilly, 1979; Yeager, 1978). A system based on mistrust of hierarchy can result in individuals distancing themselves (disengagement behavior), an unwillingness to cooperate, or result in individuals banding together in opposition (Kumar, Van Dissel, & Bielli, 1998; Lu, 2008).

*3.2.2. Bottom-up forces*

Based on the course of action taken at the organizational level, individuals will make an initial assessment, and determine if the policies, rules, and procedures are trustworthy. In this regard, individuals have agency: “actors are knowledgeable agents with a capacity to reflect and act in ways other than those prescribed by taken for granted social rules and technological artifacts” (Garud, Hardy, & Maguire, 2007, p.961). Structures imposed on individuals are not static; instead, they are something created and recreated by individuals through a process of enactment, which is temporal in nature. The norms and values of an individual, or a group of individuals, may align with the values of the organization. If perceived as being credible, trust is likely to be reinforced and established among individuals.

When misalignment occurs between the expectations, norms, and values instigated by upper levels, and those lower down, it can give rise to conflict, disagreement, and mistrust (Jackson & Fearon, 2014). Individuals may break away from existing organizational structures, and start to reframe or re-categorize their own rules, norms, and practices as a way of re-establishing trust processes (Battilana, 2006; de Lange, 2019; Luo, Chen, & Chen, 2021). As stated by Lumineau and Schilke (2018, p.243), “with existing organizational structures no longer fully determining the trust formation process, the individual may thus begin to break with existing organizational procedures and – perhaps unknowingly – start to develop a new pattern of trust formation routines.” Reframing, for instance, can be instigated by new (untested) practices, which bring uncertainty, rules, or technology introduced without consent, or when organizational managers are ignorant to different views and opinions. These individual norms may be espoused and taken up by other individuals who share a strong sense of support and belonging, making support for change more pronounced.

*3.2.3. Conceptual model and research propositions*

Drawing on the embedded agency perspective, our conceptual model takes account of both top-down and bottom-up forces and the interplay between them to understand trust/mistrust in algorithmic grading. On the one hand, organizational factors can influence individuals and the trust they have towards organizational decisions. On the other hand, individual level activity can in turn influence practices and trust processes at the organizational level, leading to changes in norms, rules, and policies. Collective bottom-up action, for instance, may force an organization to take corrective action by changing its practices or to realign its values and norms to reduce instances of mistrust. Figure 1 summarizes the conceptual model used for this study. Accordingly, the following two propositions are put forward:

Proposition 1: Top-down forces operating at the organizational level can influence lower-level activity, stimulating AI-related mistrust.

Proposition 2: Bottom-up forces operating at the individual level can influence higher-level activity, stimulating organizational readjustments to reduce AI-related mistrust.

Context

Organizational level

*Organizational norms, decisions, policies, and procedures*

**Trust/Mistrust in AI**

*Readjustment of norms, decisions, policies, and procedures*

Individuals

Figure 1. Conceptual model

**4. Methodology**

*4.1. Research setting*

This study draws on the case of the International Baccalaureate Organization (IBO) and examines the case of the algorithm used to grade IB exams in 2020, the year of the COVID-19 pandemic (Simonite, 2020). IBO, which was established in 1968, is an organization that has over 1,900,000 students across 5,000 schools globally in approximately 159 countries (IB, 2022a). The organization offers a 2-year diploma program aimed at students aged between 16-19 and is recognized internationally by many leading universities (Wright & Lee, 2020). According to its website, “IB programs develop inquiring, knowledgeable and caring young people who are motivated to succeed. The IB gives students distinct advantages by building their critical thinking skills, nurturing their curiosity, and their ability to solve complex problems” (IB, 2022b). Programs are offered in different languages, including English, Spanish and French, and can operate effectively with national curricula.

In 2020, the diploma was taken by over 170,000 students (Simonite, 2020). Prior to the pandemic, assessment normally consisted of coursework and final exams, which are administrated by IBO. Like other grading systems, e.g. UK A-Level grading system, before students sit exams, teachers provide predicted grades, which allow universities to offer a place, provided the student meets their predicted grades. Schools also send student coursework samples to externals to ensure grades are marked consistently and fairly.

The COVID-19 pandemic has caused unprecedented disruption to the educational sector worldwide, contributing to the redesign of learning and assessment. In 2020, faced with a tough decision, IBO had to decide how best to predict student grades due to the problems caused by the pandemic. As IB students could not physically sit their exams, working with an unnamed educational provider, an algorithmic model was constructed to assist with calculating student final grades (Edwards, 2021; Mouille, 2020). The algorithm worked by using both historical and current data, including student coursework marks, teacher predicted grades and school context. As noted by Owen (2020, para.8) the school context was “not based on previous cohorts’ performance, but instead, the relationship between predicted grade accuracy, performance in coursework versus examination components and final outcomes.” Given the approach used by IBO, this raises interesting questions regarding the appropriateness of using algorithms for grading purposes. The context-specific adoption of the IBO algorithmic grading in 2020 makes this a suitable case for investigating the extent of trust/mistrust of AI.

*4.2. Research design and data source*

The study is based on the constructionist research paradigmand examines stakeholder responses (e.g. IBO, students, teachers, parents) to the use of algorithmic grading adopted in IB assessment in 2020, the year of the COVID-19 pandemic. The qualitative approach was deemed suitable for the study. In particular, this is used as a means to explore human interaction and to identify patterns of behavior (Bryman, 2004) among those impacted by the algorithm, both as individuals and as a collective entity.

The primary data source for this study was an online petition campaign initiated by an IB student whose grades were downgraded by the adopted algorithm; the petition was signed by other IB impacted stakeholders, including students, parents, and teachers. Moreover, the selection of the online dataset was crucial. This needed to meet two core criteria: first, it had to show a sufficient amount of activity that would justify an in-depth investigation of their interactions, and second, it had to engage AI stakeholders in a collective activity. The petition was signed by more than 25,000 individuals around the world. Comments consisted of more than 45,000 words. For this study, we explored and analyzed the posts added by the participants of the online petition for the purpose of developing an understanding of the nature of reactions to the algorithmic grading held among those directly affected and the impact that these had. The petition was called ‘Justice for May 2020 IB Graduates - Build a Better Future! #IBSCANDAL.’

To support arguments put forward, the study also draws on relevant published material, including newspaper articles, education magazines, press releases, as well as other online resources. We list these in Appendix A.

*4.3. Analytical approach*

Adhering to the guidelines of the thematic analysis approach (Braun & Clarke, 2006), the second author inductively analyzed the data. During this process, the six phases of thematic analysis were followed: familiarization with data, initial code generation, theme search, theme review, theme definition and naming, and writing-up (Braun & Clarke, 2006).

The initial phase (phase 1) involved an iterative reading of all the data. This involved collecting all relevant documentation and online petition posts (Appendix A), which helped to develop familiarization of the case whilst enabling us to assess whether the data we had available could develop a coherent story.

Where this was not the case, additional secondary data was gathered (e.g. how IBO responded to the petition and wider concerns). During phase 2, open coding led to the identification of five recurring codes: motives for taking part in the petition, personal circumstances, impact of AI grading, including university entry, emotional state, and demand for change.

In phase 3, guided by our dataset and influenced by the overarching research question of the study, further categorization of the data took place. This centered around the impact that the adopted AI formula had on trust/mistrust by those affected. However, whilst trust/mistrust remained the focal point of analysis, we were open to other thematic categories.

The emerging categories were thereafter reviewed and verified by the first author (Phase 4), and when agreement was reached, they were named (Phase 5) as endorsing collective action, mistrust towards IBO, mistrust towards the AI formula, and reinforcing trust towards teachers and schools. Throughout the analysis process, we adopted a “theory-driven” approach (Braun & Clarke, 2006) whereby we used our specific research question, theoretical tenants of embedded agency theory, and the proposed development in Figure 1.

**5. Findings**

By applying an embedded agency perspective, we base the findings around two types of cross-level influences: top-down (organizationindividuals) and bottom-up (individualsorganization). Within these two types of cross-level effects, four core sub-themes are identified: mistrust towards IBO (5.1), mistrust towards the AI formula (5.2), which are grouped under top-down and endorsing collective action (5.3), and reinforcing trust towards teachers and schools (5.4), which are coded under bottom-up.

*5.1. Top-down forces*

Final grades play a major part in determining university admissions, as well as direct routes into employment and apprenticeships. When student final grades were released in July 2020, students discovered that their score deviated sharply from that predicted, which, for many, affected their chances of securing their preferred university choice. It later emerged that the decline in grades was attributed to an algorithm. The algorithm was implemented within the core IB’s service. A key theme that surfaced was stakeholder *mistrust towards IBO*, particularly because of its decision to adopt algorithmic grading instead of relying on teachers’ grades and its lack of transparent communication. The use of an algorithm differed considerably from established norms of sitting exams and human involvement with grading. A common sentiment that surfaced was the perception that IBO placed too much trust in an algorithm to determine scores.

*“IBO trusts more its own secret and random grading algorithm than those teachers who*

*are trained by IBO*[[1]](#footnote-1)*.”* (RY, Canada)

Prior to the release of final grades, respondents felt that there was a lack of clear policy and communication from IBO with regards to how grades will be calculated.

*“I am signing this petition because my daughter was denied her IB certification based on a formula that wasn't communicated to all.”* (JT, United States)

A further theme that emerged was stakeholder *mistrust towards the AI formula*. Mistrust was raised about those behind its development and the nature of this algorithm. There was the perception that the algorithm was biased and lacked explainability. Due to excessive trust being placed in the hands of developers, it was difficult for those affected to accept that the algorithm acted in a neutral and fair manner.

*“Having an AI algorithm (that has the developers' bias coded into it) determine the future education potential of an actual human being is insane.”* (KE, United States)

Concerns were raised specifically in relation to the formula used to predict students’ grades. A key source of skepticism was the lack of transparency in relation to how the algorithm worked. Many teachers, parents, and students were puzzled as to how the formula calculated the grades, with some questioning by what means the algorithm was designed and tested.

*“The algorithm used does not make sense and is not transparent.”* (KK, Singapore)

The approach used to calculate final grades differed considerably from customary norms and practices, where in previous years, final exams were graded and moderated by human markers affiliated with IBO. Although information as to the exact workings of the algorithm was not made publicly available, it became apparent that the grading calculation was based on a student’s assignment scores, teacher predicted grades, as well as school context.

While it was acknowledged that school context was not based on the performance of previous cohorts, there was the perception that school context may downgrade a good student who was from a school that underachieved.

*“I feel that the results are unfair, quite random, defy logic of any sort, and not transparent. Also comparing the results across schools it seems to be elitist - where selective high-end schools have favorable bias while inclusive schools with a range of students have suffered quite a bit. All the results suggest it is a bell curve around the historical average of the school with total disregard to any individual student’s capability or track record. Feel like IB[O] is showing how it does not want to hold its own principles - preach but not follow...”* (PG, India)

Despite the intention of the algorithm to create a tailored equation for each school and student, respondents expressed the inhuman nature of the algorithm, particularly its impersonal character.

*“These students that have already had to suffer through the COVID situation and have not been able to sit their exams have been punished for this by what seems to be an unjustified and impersonal algorithm!!”*(JL, Switzerland)

A legitimacy gap occurred between how individuals believed IBO should behave, based on their global credibility and high education quality standards, and the actual actions taken by the firm. This discrepancy destabilized legitimacy, promoting values of deception and mistrust at lower levels. Referring to Figure 1, top-down forces at the organizational level influenced the formation of mistrust among lower-level actors, particularly students, teachers, and parents.

*“Obviously IB[O] cared more about displaying Firm Harsh grader image than projecting their students actual academic level which is a complete shame to their reputation. Reckon class 2020 reflected how fake IB[O] objective is and how unworthy they were of our trust.”* (OA, Germany)

*“I, like many other students feel like I've been cheated by the system we've put so much trust into for the last two years. We've gone through their program, IB[O] made us swallow their BS about their magic formula which would somehow calculate our grades on the basis of some expert analysts and their vast previous data and experience--what we get from all of that, and what it really translates into, at the end of the day is a slap in the face.”* (CG, United States)

*“There can’t be such a huge difference between predicted and actual… don’t trust IB schools in India. It’s a sham.”* (RS, India)

*5.2. Bottom-up forces*

A further theme that emerged from the data analysis was *endorsing collective action*. Many students, because of their grades being downgraded, lost their university place and blamed IBO for playing with their future. Some schools, on behalf of students and parents, lodged complaints to IBO. As noted by Civinini (2020, para.4), “out of the 3,020 schools receiving results in the May 2020 session, around 700 schools submitted a review request on behalf of their students.”

The lack of response from IBO to address this urgent matter resulted in respondents feeling that the firm was refusing to be accountable for its actions. There was a sense that IBO came short of refusing to apologize for their mistakes, and instead, were attempting to defend their actions as a way of protecting their reputation.

*“*[Referring to the name of a 6th form college] *have admitted examination 'mistakes' yet the IBO refuse to even acknowledge. This is scandalous. These children's futures are being taken from them with a slip of what is clearly a dodgy formula. Hugely saddened.”* (JE, United Kingdom)

*“Disappointed with IB[O]’s lack of response after receiving overwhelming complaints on their Algorithm. Not a responsible organization you can trust.”* (JL, Hong Kong)

What followed was an “international outcry”, with the aim to put collective pressure on IBO to reconsider its policies, decisions, and procedures. An online petition was set up and signed by over 25,000 respondents. The dataset reflected the international background of those affected by the IB AI grading. Those who signed the petition were students, parents, as well as teachers who took the opportunity to express their feelings, opinions, and emotions as a result of what they referred to as the unfairness caused by the IB grading algorithm. Those who signed the petition clearly endorsed the attempt made by the initiator of the petition to put pressure on IBO to reconsider its grading policy. By so doing, there was an effort to promote a strong voice and collective action, forcing IBO to reverse its decision and carry out a re-evaluation of the grades.

*“I’m an IB student who did NOT get their IB diploma! I scored 23 points which should not be the case since 1) I scored high on my IA’s (according to the IB rubrics my teachers used) and 2) I scored significantly lower than my predicted score. Let’s make a change and make the IB[O] realize that THIS IS NOT OKAY!!!”* (JR, United States)

Moreover, the online petition itself that was set up illustrates further endorsement of collective action, showing enthusiasm that this initiative has started:

*“Thank you very much, @AliZagmout, to give a voice to all these students and their parents who gave everything to make a good IB only to be treated with such injustice at the end that will take them many chances and shut doors that would have been open without Corona.”* (CR-R, Germany)

Collective voice and action were important to bring “shame” on IBO and to acknowledge that students “deserve better” than the service and approach provided. Students and parents demanded greater communication and transparency as to how the algorithmworked, and why it was used to determine student grades.

*“I'm an IB teacher and the students deserve better than this.”* (FC, France)

*“I am signing because students and IB teachers deserve to know why this happened!”*(LE, United States)

*“The IB[O]'s business model depends on their reputation, nothing else. They will not change for individuals, only collective action to shame them into putting this right.”*(HC-T, United Kingdom)

Whilst mistrust towards both the grading algorithm and IBO was clearly stated in the messages posted, there was also evidence emphasizing the need for greater trust to be placed in teachers and schools. Unlike the impersonal algorithm, teachers and schools were in a better position to carry out fair assessment due to their close contact with students; as a result, these were seen by students and their parents as trustworthy, and their views trusted.

*“A teacher's prediction on the other hand takes into consideration the individual student as a whole. This, therefore, means that the teacher's prediction is a more accurate representation of a candidate's capabilities.”*(IB, Norway)

*“Many students deserve the higher grade as they have worked hard for it and their teachers should be trusted!”*(KM-C, United Kingdom)

The above posts shared in the online campaign show individuals’ motives for signing the petition but also an acknowledgment that by having a collective voice, the affected individuals could make a stronger impact on influencing IBO to reverse its grading decisions. Through a collective voice, participants could openly express their concerns and sense of mistrust that they developed towards IBO and the AI formula. By openly showing trust towards their teachers and schools, petition participants were asking for a U-turn in the decision made by IBO so that it does not rely on the algorithm.

Referring to Figure 1, emergent forms of action and behavior from lower levels can influence changes at the higher level, including trust processes. While the IBO initially defended its use of an algorithm to determine grades, in the face of increased scrutiny, it underwent readjustment of norms, decisions, policies, and procedures to re-establish trust. Following a strong, collective voice, rather than placing unbridled trust in the use of algorithms for grade predictions, it was decided that greater trust would be *reinforced towards teachers and schools*. More specifically, a new procedure (“additional review service”) was established, allowing schools to submit supplementary evidence for review if results were misaligned with anticipated student performance (IB, 2020a). The aim of this data was to help make sense of what went wrong and to prevent the same mistakes from happening again in the future.

IBO revised its policy that when a student’s coursework grade (marked internally) was within one mark of the teacher’s predicted grade, then this mark would be revised and would become the final grade (Civinini, 2020; Jack, 2020). This would ensure that students to who this policy applied would not have their grades downgraded. The aim of this policy was, as reported by IBO, to ensure “validity” and “fairness”. As noted in a statement by IBO: “*The award of the revised final grade is based on these two data points of internal assessment and predicted grades to ensure the validity and fairness of the final grades*” (IB, 2020a). Part of IBO’s trust-building process was to reaffirm its position that trust must be earned, and not given, and it would strive to maintain this long-lasting bond of trust that it worked hard to achieve: *“We have earned your trust for 50 years and in this unprecedented year we continue to serve our community”* (IB, 2020a). Going forward, through communication updates and guidelines, stress was given to norms and values reinforcing equality, transparency, and fairness (IB, 2021).

It is our position that a fuller understanding of AI-trust/mistrust requires an appreciation of its multilayered nature, cross-level effects, as well as context. Rather than considering trust at just one level, or focusing on a specific target e.g., organization, group, or individual, our findings bring various levels together. As demonstrated, trust/mistrust formation in algorithmic grading is embedded within an intricate web of relationships across multiple levels (e.g., IBO, schools, teachers, parents, and students). Instead of simply illustrating how top-down forces directed by organizational practices can, in the case of the IBO, exert a constraining effect, our findings demonstrate the salient and recursive nature of bottom-up forces (e.g., schools, students, parents, teachers), both individually and at the group level, which can influence trust processes. In other words, trust is both top-down constrained by organizational (hierarchical) mechanisms, as well as influenced by bottom-up emergent processes as individuals come together to form collective attitudes towards trust/mistrust. Additionally, trust/mistrust in AI is not formulated in a vacuum and is shaped, both directly and indirectly, by its situational context. Events, e.g., the COVID-19 pandemic, can seek to destabilize trust and it is important that contextual factors are considered carefully when deploying AI.

**6.** **Discussion**

Referring to the left-hand side of Figure 1, the case shows that top-down forces, such as organizational norms, decisions, policies, and procedures may influence trust. Since its foundation in 1968, a key norm of IBO is not only to deliver high-quality teaching but also to offer “trusted methodologies and experience in assessment” (IB, 2020b). Historically, the organization has prided itself in earning the trust of the community. In the case of its decision to use an algorithm for assessment, this, as perceived by many, conflicted with IBO’s customary policies and procedures of providing robust assessment methods.

The use of an algorithm to predict grades reinforced a constraining environment at lower levels. The situation that emerged within a very short space of time, at both the individual and group levels, was one of frustration, anger, and mistrust. Students, teachers, and parents united in opposition, forming collective action and raising fundamental concerns about the fairness, bias, and transparency of the algorithm. These findings are similar to other studies (e.g. Fitzpatrick, Friend, Costley, 2004; Jackson & Panteli, 2021; Nam et al., 2020), who have shown that when mistrust arises, it can foster an environment of dissatisfaction or resentment.

When communication from upper levels lacks transparency, and the intended purpose of an initiative does not meet the expectations of individuals, it can nurture an environment of skepticism and mistrust (Huvila, 2017; Jackson & Fearon, 2014). Saunders and Thornhill (2004, p.10) note “the quality of information and the consistency between management’s stated strategy and the realities is also likely to impact upon trust and mistrust.” In the case of IB, details regarding the grading process were not immediately clear. Students expected a system based on openness and fairness, but these expectations were not met. Polyakov (2020) outlines six key reasons why AI can lead to mistrust: safety, security/robustness, privacy, transparency and fairness, ethics and sustainability, and accountability. Clearly, in the IB case, mistrust was largely due to lack of transparency and fairness, especially at a time when students’ education was hugely disrupted, with homeschooling and lack of face-to-face teaching being the norm during this period. Despite normally being regarded as a trustworthy organization, trust quickly eroded. Based on an analysis of the findings, we can offer support for proposition 1.

One key explanation as to why stakeholder trust faded quickly may be due to many educational systems across the globe following a highly institutionalized model and centralized governance structure. The COVID-19 pandemic, and the sudden move to, in many cases, forced online learning conflicted with traditional institutional norms of formal grading procedures and exam standards, which typically involve human elements. In the educational sector, for example, a large component of marking is placed in the hands of internal and external markers, which often entails first marking and stringent moderation procedures. Long-established norms and taken-for-granted rules of conducting assessment using in-person exam rooms, invigilators, and other prescribed standards, e.g. assessment format, are not only aimed at upholding exam integrity by preventing malpractice, but also ensuring equal opportunity. These norms and practices can reinforce certain stakeholder expectations, norms, and values in relation to how assessments should be conducted and what is perceived as legitimate. The use of what was considered by many students as an untested and novel approach to grading conflicted with students’ traditional norms and expectations, which involved human marking, proven assessment requirements, and impartial exam procedures. Unfortunately, in the case of the COVID-19 pandemic, with the move to online modes of teaching and assessment, IBO experienced challenges in maintaining customary assessment standards.

The study also illustrates how emergent forms of action and behavior at lower levels can influence changes at the higher (organizational) level, as shown on the right-hand side of Figure 1. In times of crisis, such as a pandemic, social structures may be more inclined to experience ruptures or breakdowns (Risjord, 2014). Stakeholders, including students, teachers, and parents, interpreted algorithmic grading and the actions taken by IBO as highly disruptive. As opposition to the algorithm grew, IBO decided to ease the appeal process and, in some cases, grades were adjusted based on teachers’ recommendations. Abdelnour et al. (2017 p.1776) note “actors have much greater leeway to interpret rules and enact institutional patterns and relationships than previously assumed in institutional theory.” The organization reflected on and learned from the challenges that they encountered using an algorithm for grading purposes. In so doing, it changed its processes and procedures for calculating and awarding final grades, as well as introducing mechanisms for ensuring validity and fairness in meeting the future needs of the organization and students. Trust was repositioned from the use of an algorithm to being centered on teachers. Based on the evidence we can substantiate proposition 2.

Although the study illustrates that bottom-up activity contributed to top-down (organization) change, one important question is: What higher-level factors may have contributed to the readjustment of norms, policies, and procedures? Since strong opposition to the actions of an organization can run the risk of endangering its success and mistrust can have a major negative impact on a firm, businesses may engage in strategies that attempt to reduce institutional pressures and reputational harm (Zaharopoulos & Kwok, 2017). Drawing on agency theory, from an economics perspective, response strategies used by firms may be based on their self-serving motives when dealing with a crisis situation, such as a pandemic. Since financial incentives, i.e. management bonuses, promotion, and individual risk are linked to the health and performance of an organization, managers responsible for initiating change initiatives may have the propensity to act in their own (self-serving) interests (Abrahamson & Park, 1994). In the case of IB, to avoid a potential loss of income through students withdrawing from programs, organizational managers may be more inclined to conform to collective pressure. As numerous stakeholders (e.g. community members, teachers, students, parents) can directly or indirectly exert influence on a firm and the firm can be influenced by the activities of relevant stakeholders, it is important that the firm attempts to satisfy their needs and desires (Hooghiemstra, 2000). If a firm is not seen as acting in a manner that is perceived by society members as legitimate or operating in societal group interests, it can greatly erode continued trust in the firm (Pfeffer & Salancik, 1978). Consequently, this may lead to loss of organizational support, reduced demand for educational services, or possible legal action.

*6.1. Theoretical implications*

This paper raises several important theoretical implications. Although a large body of work has examined trust in IS research, fewer studies have examined trust in the context of AI, and more specifically, algorithmic grading. Studies that have examined trust have largely focused on the individual level. This study advances our understanding of trust in relation to AI. Although trust does emulate at the individual level, other levels can also influence, both directly and indirectly, trust/mistrust development. Our position is that trust in AI is best considered a multi-level dynamic phenomenon. We advance the AI-human dynamism perspective by examining the lively interplay between human and technical elements, and how, looking at the intertwinement of multiple levels including top-down and bottom-up forces, can influence trust/mistrust in algorithmic grading across time and space.

The study offers support for an embedded agency perspective on the study of trust and AI by moving beyond traditional notions of conformity and stability and putting agency at the center of attention. Rather than focusing exclusively on how organizational structure can influence or constrain lower-level activity, based on an interpretation of findings, our study illustrates how agents can come to intentionally change social structures. Agents have the ability to act, they have power, as well as the capacity to bring about change. Furthermore, the situation in which the AI is developed and used may itself have an influence on the degree of trust/mistrust. Therefore, these influences should be part of a co-creative effort rather than at separate stages, as was the case with the IB example.

Having devised a conceptual model based on embedded agency from institutional theory, one important question is: How does our proposed model advance our understanding of institutional theory? It is important to note that given the diverse and fragmented nature of work utilizing institutional theory, we do not claim to build or offer an all-embracing theoretical model that captures the complex nature of institutions. Rather the focus of our theory development is on the area of trust/mistrust surrounding algorithmic grading.

Although much effort within institutional theory has been conducted to understand how top-down forces can constrain lower levels, still much effort remains to explore the activities occurring at the bottom, and how this can influence higher institutional norms and practices (Cardinale, 2018; Lumineau & Schilke, 2018). By developing a holistic multi-layered perspective, we have attempted to understand further how the presence of top-down and bottom-up activity can exist concurrently, as well as exhibiting cross-level effects. Our model, and subsequent findings, illustrates how top-down actions can exert a constraining influence at the individual level, thus reinforcing mistrust, but also how emergent lower-level activity can play a role in restructuring institutional contexts. An important point raised is that trust/mistrust should not be solely treated as the property of any one level, rather trust/mistrust comes into existence through the interconnectedness and interactivity as different levels come together, within a changing context.

Being mindful of the need to move beyond traditional notions of collectivity, conformity, and stability within institutional theory (Haack, Sieweke, Wessel, 2019), a further contribution of this study is an attempt to put heterogeneity, change, and agency at the center of attention. Instead of considering bottom-up activity merely as collective creations that exist initially, we advocate the importance of considering the varied actors at play (in our case students, teachers, parents, and community members), and how collective action is something in the making. Therefore, it is best to think of collective activity as soft-assembled—a temporal (suspended) property of group action that is ebbing and flowing as individuals come together as they respond to the situational context. Institutions are not fixed but are continuously being created and recreated by individuals as the respond to the environment around them. Additionally, we shed light on how the interactions occurring between individuals can give rise to emergent action and outcomes. Our study, for instance, elucidates the transformational role that agents operating at the bottom can bring in reforming institutional norms and changes in the use and adoption of AI.

*6.2. Practical implications*

Based on a reflection of the findings, the paper raises several important implications for a diverse group of stakeholders both independently and collectively. Independently each of the stakeholders groups can exert their influence based on their own interests and expertise; collectively there are opportunities for co-creation of AI contributing to the wider AI acceptance.

*Developers:* there is a need to move beyond the black-box nature of AI and place greater emphasis on explainability and understandability. This is not to say that those impacted by AI need to know the technical and mathematical intricacies of how the system works, but, at the very least, there should be clear justification as to why the AI is being introduced, and its intended purpose presented in a way that is easy to comprehend. Developers, in layman terms, should be able to explain why an algorithm reached a particular decision, and where and what type of data is being collected and analyzed in the process. By explaining AI in simple terms and endorsing understandability, this will help to ensure transparency in the design and deployment of AI-based systems and assist in the initial development of trust.

Additionally, as indicated in our study, the development of algorithms that are deemed bias can have significant implications for the establishment of trust. Thus, careful attention needs to be given to ensure that, where possible, bias is removed in the training of AI models and algorithms. Extra care needs to be taken when programming for AI systems—ensuring that any potential programmer bias is eliminated or reduced, as this can significantly erode trust. This can be achieved by providing adequate training programs in implicit bias avoidance. Training should focus on understanding types of algorithmic bias and potential sources of bias, as well as the ways in which one’s own bias can be overcome. AI ethical procedures and measures should be put in place to ensure professional accountability mechanisms, self-governance practices, as well as adherence to appropriate ethical codes of conduct and policies. Furthermore, it is important for developers to comprehend the needs of those who will be impacted, both directly and indirectly, by AI. Needs assessment should be based on the principles of diversity and inclusiveness—all stakeholders need to be equally represented and impartiality needs to be enforced to bolster up trust in AI. This may help in the design and development of AI that is ethical and fair.

*Organizational managers:* managers must be open and sensitive to issues that emerge throughout the AI implementation process. Evaluation, monitoring, and feedback mechanisms should be put in place to flag any potential issues allowing for early warning sign detection. For warning detection to be effective, stakeholders involved in AI need to work together in synergistic ways, raising any problems, challenges, and obstacles that may give rise to mistrust. Clear milestones and checkpoints throughout the AI project need to be established. Contingency planning for AI needs to be given priority and integrated with the overall strategic planning process. Managers need to invest and put in place effective governance procedures and processes to ensure that trust and ethical standards are met and adhered to.

Mistrust awareness should become part and parcel of the AI lifecycle. If mistrust arises, managers should be responsive and commit dedicated time and resources, making sure that mistrust is lessened, or proactively managed. This should involve considering the views and opinions of those impacted by AI and attempt to address any concerns raised. If necessary and feasible, managers should adjust the features of the AI to accommodate the needs of affected stakeholders. If organizational managers are unable to foster an orientation of trust to support AI, then managers should consider very carefully if the AI solution or implementation approach is the correct choice for the needs of the organization.

Often unsatisfactory outcomes with AI can be attributed to a poor fit between the existing culture and failure to nurture a conducive culture to support AI. It is important to note that trust in AI is not automatically built suddenly, nor can mistrust attributed to AI be eradicated overnight. If mistrust cannot be controlled in the short term, then managers should be realistic. A more subtle/incremental approach with minor AI additions and adjustments rather than radical change, as well as using AI as a supplement rather than an alternative to existing systems and processes, where possible, should be adopted. If it is anticipated that the AI system will bring major challenges, managers should be prepared to face a lot of opposition, as well as expect to spend substantial time and resources to correct the situation. Alternatively, managers should question if an AI solution is truly what the organization needs.

*Policy makers:* the development of AI-related policies should focus on, and be guided by, a human centric approach—policy practices need to center on how AI can complement and enhance the human experience. In the case of AI for educational purposes, a clear connection needs to be established between AI and its benefits for education. This should focus on the principles of equitability and inclusiveness, by ensuring that students from different socio-economic backgrounds, as well as different abilities, are equally and fairly represented. Policies must set out established norms, standards, best practices, and ethical guidelines that treat AI as a public good—one that empowers, not threatens, educators and students, nor goes against their ethical norms, values, beliefs, and standards.

It is important to be mindful that no one-size-fits-all policy will work, and consideration needs to be given to the context of use, and what implications this has for stakeholder trust in AI. For instance, while our study found transparency, explainability, accountability, and fairness, to be key elements in the formation of AI trust or mistrust in algorithmic grading, other elements e.g., security, privacy, safety, reliability, may be equally as important. Therefore, in terms of policy formulation and realization, it is important to be considerate of the characteristics, the industry, and the anticipated impact and outcomes of AI to be deployed.

The development of AI, and how it is to be used, should not solely reside in the hands of a single entity, rather it requires major collaboration within and across academia, industry, and government. Collaborators e.g., educators, technologists, AI engineers, and other social scientists should bring with them interdisciplinary expertise to assist in policy advice around the design, development, and use of AI initiatives. Policy should not be set in stone, and open to ongoing review and modification. Approaches to AI governance and planning at the organizational level should align with AI and policies both at local and national levels.

*School community:* a common conception is that mistrust arises due to the threat of AI eliminating the workforce or making existing job functions obsolete. Rather than AI being perceived as a way of reducing teachers or eliminating some of the core functions that educators perform, they should be promoted as empowering. Teachers need to feel that their abilities and contributions are important and valued. Manual grading can be a grueling process that can often consist of various assessment types and points throughout the academic year. If teacher grades are simply going to be disregarded and replaced by an algorithm at short notice or without adequate consultation, it is not surprising that teachers will feel undervalued and mistrustful. A key lesson learned from our study is that ownership of grading, where possible, should be placed in the hands of teachers. Educators should not be ringfenced into following a particular pedagogical practice e.g., algorithmic grading, but should be given flexibility to overwrite decisions made by algorithms and take action based on their own intuitions.

AI-tools need to be considered as effective and beneficial from a student perspective. Given the fear that algorithms may not award grades fairly, as well as the perceived novelty associated with AI for educational purposes, a challenge is to facilitate the emergence of the right environment to enable students to move to the comfort zone without dragging them into the fear zone. Students should be provided with an overview of the basic features of the AI being proposed upfront as this would influence their ability and willingness to accept it. The aim should not be to provide too much detailed technical knowledge. There should be enough information (procedural transparency) to allow for an uncomplicated overview of the process and data sources used by the algorithm to reach a specific outcome—one that allows students to perceive the result reached as unbiased, valid, and fair.

*6.3. Limitations and future research*

The findings of the study are based on an investigation of a single case of the IB grading algorithm. It cannot be claimed that the findings, and their implications, are generalizable to other types of institutions. Further studies should seek to examine the use of algorithms for grading purposes in other settings. Moreover, the focus of the study has been on algorithmic grading. Other forms of AI may embody particular patterns of trust/mistrust; this was not investigated in this study, and therefore, there is potential for further research. Further research should also seek to extend the application of our theoretical model in the case of algorithmic grading during ‘normal’ school times, as well as in other AI applications. Using sentiment analysis or a survey, additional research could assess stakeholder trust towards IBO in the years following the algorithmic grading fiasco.

An interesting avenue of future research would be to understand further the key different types and specific characteristics of trust in the context of AI. Although this study largely considered the organizational, group, and individual levels, other levels could have been explored. Further studies, in addition to the individual, group, and organizational levels, could examine in greater detail the government, right up to the society level, and how these levels influence trust/mistrust formation in AI. Another limitation of this study is that it did not explore specific situational factors (e.g. gender, age, culture, among others), and how these factors influence trust/mistrust in algorithmic grading. Further studies could explore these factors by using a mixed-methods approach, combining both qualitative and quantitative methods.

**7. Conclusion**

This paper set out to investigate trust/mistrust formation in the context of algorithmic grading by drawing on the IB grading fiasco in summer 2020. While the study of trust and computerized applications represents a growing and important area of inquiry, research on algorithmic grading and issues pertaining to trust/mistrust are still at a nascent stage, and much remains to explore these issues in greater depth. Comprehending how these areas come together is an important endeavor and may reveal unique insights not previously known or explored. As a way of advancing our understanding of trust/mistrust concerning the adoption of algorithms, as well as other forms of AI, robotics, and machine learning, there is an increased need for studies to investigate further its dynamic and multi-faceted nature. As illustrated in this study, trust/mistrust can emerge across multiple levels, and it is important to remain sensitive to issues relating to context, dynamism, and change.

**References**

Abbass, H. A. (2019). Social integration of artificial intelligence: Functions, automation allocation logic, and human-autonomy trust. *Cognitive Computation*, *11*(2), 159-171.

Abdelnour, S., Hasselbladh, H., & Kallinikos, J. (2017). Agency and institutions in organization studies. *Organization Studies*, *38*(12), 1775-1792.

Abrahamson, E., & Park, C. (1994). Concealment of negative organizational outcomes: An agency theory perspective. *The Academy of Management Journal*, *37*(5), 1302-1334.

Alter, S. (2021). Understanding artificial intelligence in the context of usage: Contributions and smartness of algorithmic capabilities in work systems. *International Journal of Information Management*, 102392.

Anderson, J., Rainie, L., & Luchsinger, A. (2018, December 10). *Artificial intelligence and the future of humans*. Pew Research Center.  https://www.pewresearch.org/internet/2018/12/10/artificial-intelligence-and-the-future-of-humans/

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, *22*(6), 1-7.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Bates, T., Cobo, C., Mariño, O., & Wheeler, S. (2020). Can artificial intelligence transform higher education? *International Journal of Educational Technology in Higher Education*, *17*(42), 1 -12.

Battilana, J. (2006). Agency and institutions: The enabling role of individuals’ social position. *Organization*, *13*(5), 653-676.

Bitkina, O. V., Jeong, H., Lee, B. C., Park, J., Park, J., & Kim, H. K. (2020). Perceived trust in artificial intelligence technologies: A preliminary study. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *30*(4), 282-290.

Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. *International Conference on Human-Computer Interaction* (pp. 3-20). Springer, Cham.

Bogina, V., Hartman, A., Kuflik, T., & Shulner-Tal, A. (2021). Educating software and AI stakeholders about algorithmic fairness, accountability, transparency, and ethics. *International Journal of Artificial Intelligence in Education*, 1-26.

Bonderud, D. (2019, August 27). *Artificial intelligence, authentic impact: How educational AI is making the grade.* EdTech. https://edtechmagazine.com/k12/article/2019/08/artificial-intelligence-authentic-impact-how-educational-ai-making-grade-perfcon

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Bryman, A. (2004) *Social research methods* (2nd ed.). Oxford University Press.

Cannon, J. (2019, December 4). *Report shows consumers don’t trust artificial intelligence.* Fintech News. https://www.fintechnews.org/report-shows-consumers-dont-trust-artificial-intelligence/

Cardinale, I. (2018). Beyond constraining and enabling: Toward new microfoundations for institutional theory. *Academy of Management Review*, 43(1), 132-155.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons.

Chandra, S., Shirish, A., & Srivastava, S. (2021). Examining user engagement in AI-based interactional tools: Valorizing the essence of human-beings. *AOM 2021: 81st Annual Meeting of the Academy of Management. "Bringing the Manager Back in Management"*, Jul 2021, Virtual, United States.

Chatterjee, S., & Bhattacharjee, K. K. (2020). Adoption of artificial intelligence in higher education: A quantitative analysis using structural equation modelling. *Education and Information Technologies*, *25*(5), 3443-3463.

Daugherty, P. R., Wilson, H. J., & Chowdhury, R. (2019). Using artificial intelligence to promote diversity. *MIT Sloan Management Review*, *60*(2), 1.

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, *96*(1), 108-116.

de Lange, D. E. (2019). A paradox of embedded agency: Sustainable investors boundary bridging to emerging fields. *Journal of Cleaner Production*, *226*, 50-63.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*(2), 147-160.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Emergen Research (2022, February 21). *Artificial Intelligence in the Education Sector Market Size, Share, Growth, Trend and Forecast Till 2027*. https://www.einnews.com/pr\_news/563665544/artificial-intelligence-in-the-education-sector-market-size-share-growth-trend-and-forecast-till-2027

Erez, M., & Gati, E. (2004). A dynamic, multi-level model of culture: From the micro level of the individual to the macro level of a global culture. *Applied Psychology: An International Review*, *53*(4), 583-598.

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust Incrementally: A Multi-layer model of trust to analyze Human-Artificial intelligence interactions. *Philosophy & Technology*, *33*(3), 523-539.

Fitzpatrick, M., Friend, L., & Costley, C. (2004). Dissatisfaction and distrust. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior*, *17*, 117-129.

Fulmer, A. & Dirks, K. (2018). Multilevel trust: A theoretical and practical imperative. *Journal of Trust Research*, *8*(2), 137-141.

Garud, R., Hardy, C., & Maguire, S. (2007). Institutional entrepreneurship as embedded agency: An introduction to the special issue, *Organization Studies*, *28*(7), 957–969.

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine*, *1*, 100001.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Goel, A. K., & Joyner, D. A. (2017). Using AI to teach AI: Lessons from an online AI class. *AI Magazine,* *38*(2), 48-59.

Haack, P. Sieweke, J. & Wessel, L. (2019). Microfoundations of institutions. *Research in the Sociology of Organizations*, 65A, 11-40.

Hengstler, M. Enkel, E. & Duelli, S. (2016). Applied artificial intelligence and trust: The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change, 105*, 105-120.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

Hooghiemstra, R. (2000). Corporate communication and impression management – New perspectives why companies engage in corporate social reporting. *Journal of Business Ethics*, *27*(1), 55-68

Hradecky, D., Kennell, J., Cai, W., & Davidson, R. (2022). Organizational readiness to adopt artificial intelligence in the exhibition sector in Western Europe. *International Journal of Information Management*, *65*, 102497.

Huvila, I. (2017). Distrust, mistrust, untrust and information practices. *Information research*, *22*(1), 1-14.

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, *1*, 100001.

Jackson S., & Panteli N. (2021) A multi-level analysis of mistrust/trust formation in algorithmic grading. In: Dennehy D., Griva A., Pouloudi N., Dwivedi Y.K., Pappas I., Mäntymäki M. (eds) *Responsible AI and analytics for an ethical and inclusive digitized society*. I3E 2021. Lecture Notes in Computer Science, vol 12896. Springer, Cham.

Jackson, S. & Panteli, N. (2020). Trust in the era of artificial intelligence: A multi-layer analysis. *34th Annual British Academy of Management Conference*. Manchester, England.

Jackson, S., & Fearon, C. (2014). Exploring the role and influence of expectations in achieving VLE benefit success. *British Journal of Educational Technology*, *45*(2), 245-259.

Kaplan, J. (2016). Artificial intelligence: Think again. *Communications of the ACM*, *60*(1), 36-38.

Karandish, D, (2021, August 16). *7 ways AI can benefit education*. ECampusNews,https://www.ecampusnews.com/2021/08/16/7-ways-ai-can-benefit-education/

Kaur, S., Tandon, N., & Matharou, G. S. (2020). *Contemporary trends in education transformation using artificial intelligence*. In V. Garg., & R. Agrawal. Transforming Management Using Artificial Intelligence Techniques (pp. 89-103). CRC Press.

Kavitha, P., Moorthy, B. K., Sudharshan, P. S., & Aarthi, T. (2018). Mapping artificial intelligence and education. In *2018 International Conference on Communication, Computing and Internet of Things (IC3IoT)* (pp. 165-168).

Khare, K., Stewart, B., & Khare, A. (2018). Artificial intelligence and the student experience: An institutional perspective. *IAFOR Journal of Education*, *6*(3), 63-78.

Kholeif, A. O. R., & Jack, L. (2019). The paradox of embedded agency from a strong structuration perspective: An illustrative case study of resistance to change in budgeting processes. *Qualitative Research in Accounting & Management, 16*(1), 60-92.

Klein, K., & Kozlowski, S. (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions and new directions*. CA Jossey-Bass.

Kolbjørnsrud, V., Amico, R., & Thomas, R. J. (2016). How artificial intelligence will redefine management. *Harvard Business Review*, *2*, 1-6.

Kumar, K., H. Van Dissel., & P. Bielli. (1998). The merchant of prato--revisited: Toward a third rationality of information systems. *MIS Quarterly, 22*(2), 199-226.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2), 1-16.

Lawrence, T. B., Suddaby, R., & Leca, B. (2009). Introduction: Theorizing and studying institutional work. In T. B. Lawrence, R. Suddaby, & B. Leca (Eds.) *Institutional work: Actors and agency in institutional studies of organizations* (pp. 1-27). Cambridge University Press.

Liu, B. (2021). In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384-402.

Lu, J. L. (2008). Organizational role stress indices affecting burnout among nurses. *Journal of International Women's Studies*, *9*(3), 63-78.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson Education.

Lumineau, F., & Schilke, O. (2018). Trust development across levels of analysis: An embedded-agency perspective. *Journal of Trust Research*, *8*(2), 238-248.

Luo, J., Chen, J., & Chen, D. (2021). Coming back and giving back: Transposition, institutional actors, and the paradox of peripheral influence. *Administrative Science Quarterly*, *66*(1), 133-176.

Mayer, R. Davis, J. & Schorman, F. (1995). An integrative model of organizational trust. *Academy of Management Journal*, *20*(3), 709-734.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*(2), 340-363.

Möllering, G. (2005). The trust/control duality: An integrative perspective on positive expectations of others. *International Sociology*, *20*(3), 283-305.

Möllering, G, Gillespie, N., & Lewicki, R. (2021). Multilevel trust: Reflections, insights, and a future research agenda. In N. Gillespie, A. Fulmer, & R. Lewicki (Eds*.), Understanding trust in organizations* (pp. 361-373). Routledge.

Nalbant, K. G. (2021). The importance of artificial intelligence in education: A short review. *Journal of Review in Science and Engineering*,  1-15.

Nam, K., Baker, J., Ahmad, N., & Goo, J. (2020). Dissatisfaction, disconfirmation, and distrust: An empirical examination of value co-destruction through negative electronic word-of-mouth (eWOM). *Information Systems Frontiers*, *22*(1), 113-130.

Ologeanu-Taddei, R. & Vitari, C. (2020). Filling gaps and re-conceptualizing trust in information technology. *European Conference on Information Systems*, June 15-17, 2020, Morocco.

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust.  *IJCAI 2019 Workshop on Explainable Artificial Intelligence*, August 11, 2019, Macau, China.

Parsons, J. (2020, September 4). *Students figure out how to cheat AI grading algorithm.* Metro*.* https://metro.co.uk/2020/09/04/students-figure-out-how-to-cheat-ai- grading-algorithm-13222401/

Pfeffer, J., & Salancik, G. (1978). *The external control of organizations: A resource dependence perspective*. Harper & Row.

Plitnichenko, L (2020, May 30). *5 main roles of artificial intelligence in education.* eLearning Industry. https://elearningindustry.com/5-main-roles-artificial-intelligence-in-education

Polonski, V. (2018, January 10). *People don't trust AI--Here's how we can change that*. Scientific American. https://www.scientificamerican.com/article/people-dont-trust-ai-heres-how-we-can-change-that/

Polyakov, K. (2020, December 28). Six reasons AI does not deserve our trust. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2021/12/28/six-reasons-ai-does-not-deserve-our-trust/?sh=11299ba41659

Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1-13.

Rai, A. Constantinides, P. & Sarker, S. (2019). Editor’s comments: Next-generation digital platforms: Toward human–AI hybrids,” *MIS Quarterly.* 43(1), pp. iii-ix.

Risjord, M. (2014). Structure, agency, and improvisation. In *Rethinking the individualism-holism debate* (pp. 219-236). Springer, Cham.

Roberts, K. H., & O'Reilly III, C. A. (1979). Some correlations of communication roles in organizations. *Academy of Management Journal,* *22*(1), 42-57.

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, *26*(5), 2749-2767.

Saunders, M., & Thornhill, A. (2004). Trust and mistrust in organizations: An exploration using an organizational justice framework. *European Journal of Work and Organizational Psychology*, *13*(4), 493-515.

Schiff, D. (2021). Out of the laboratory and into the classroom: The future of artificial intelligence in education. *AI & Society*, *36*(1), 331-348.

Scott, W. R. (2004). *Institutional theory*. In G. Ritzer. Encyclopedia of Social Theory. Thousand Oaks, Sage. (pp. 408-414).

Scott, W. R. (2008). *Institutions and organizations: Ideas and interests*. Sage Publications.

Seo, M. G., & Creed, W. D. (2002). Institutional contradictions, praxis, and institutional change: A dialectical perspective. *Academy of Management Review*, *27*(2), 222-247.

Sethumadhavan, A. (2019). Trust in artificial intelligence. *Ergonomics in Design*, *27*(2), 34-34.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 1-10.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541-565.

Siau, K. & Wang, W. (2018). Building trust in artificial intelligence, machine learning and robotics. *Cutter Business Technology Journal*, *31*(2), 47-53.

Smith, H. (2020). Algorithmic bias: Should students pay the price? *AI & Society*, *35*(4), 1077-1078.

Stanton, B. & Jensen, T. (2021). *Trust and artificial intelligence*. National Institute of Standards and Technology, U.S. Department of Commerce.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 272-283.

Vincent-Lancrin, S. & van der Vlies, R. (2020). *Trustworthy artificial intelligence (AI) in education: Promises and challenges*. OECD Education Working Paper No. 218, Directorate for Education and Skills.

Wang, C., Teo, T. S., & Janssen, M. (2021). Public and private value creation using artificial intelligence: An empirical study of AI voice robot users in Chinese public sector. *International Journal of Information Management*, *61*, 102401.

Welter, F. (2012). All you need is trust? A critical review of the trust and entrepreneurship literature. *International Small Business Journal*, *30*(3), 193-212.

Whitener, E. M., Brodt, S. E., Korsgaard, M. A., & Werner, J. M. (1998). Managers as initiators of trust: An exchange relationship framework for understanding managerial trustworthy behavior. *Academy of Management Review*, *23*(3), 513-530.

Wijen, F., & Ansari, S. (2007). Overcoming inaction through collective institutional entrepreneurship: Insights from regime theory. *Organization Studies*, *28*(7), 1079-1100.

Yeager, S. J. (1978). Measurement of independent variables which affect communication: A replication of Roberts and O'Reilly. *Psychological Reports*, 43(3\_suppl), 1319-1324.

Zaharopoulus, D., & Kwok, L. (2017). Law firms' organizational impressions management strategies

on twitter. *Journal of Creative Communications*, *12*(1), 48–61.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1-27.

Zietsma, C., & Lawrence, T. B. (2010). Institutional work in the transformation of an organizational field: The interplay of boundary work and practice work. *Administrative Science Quarterly*, *55*(2), 189-221.

**APPENDIX A**

*Data sources specific to IB case*

Civinini, C. (2020, August 17). *IB offers new grades to tackle 'algorithm' controversy*. TES.https://www.tes.com/news/ib-offers-new-grades-tackle-algorithm-controversy

Edwards, C. (2021). Let the algorithm decide? *Communications of the ACM*, *64*(6), 21-22.

International Baccalaureate. (2020a). *IB update on May 2020 diploma programme and career-related programme results.* https://ibo.org/news/news-about-the-ib/update-m20-dp-cp-results

International Baccalaureate. (2020b). *Awarding May 2020 results further information*. https://ibo.org/news/news-about-the-ib/awarding-may-2020-results-further-information/

International Baccalaureate. (2021). *How is the IB ensuring fairness in May 2021?* https://www.ibo.org/globalassets/programme-information/dp/how-is-the-ib-ensuring-fairness-in-may-2021-en.pdf

International Baccalaureate. (2022a). *Facts about IB programmes and schools*. https://www.ibo.org/about-the-ib/facts-and-figures/

International Baccalaureate. (2022b). *Why the IB is different. https://www.ibo.org/benefits/why-the-ib-is-different/*

Jack, A. (2020, August 17). *International Baccalaureate to revise student grades*. Financial Times.https://www.ft.com/content/2283fc3d-92a0-4480-bace-7ca3234d7942

Justice for May 2020 IB Graduates. *Build a better future! #IBSCANDAL* [Online petition]. https://www.change.org/p/international-baccalaureate-organisation-ibo-justice-for-may-2020-ib-graduates-build-a-better-future

Mouille, Z. (2020, July 29). *IB results day: A broken algorithm which decided students’ futures*. Cherwell. https://cherwell.org/2020/07/29/ib-results-day-a-broken-algorithm-which-decided-students-futures/

Owen, J. (2020, August 19). *U-turn over International Baccalaureate results*. Independent Education Today. https://ie-today.co.uk/news/u-turn-over-international-baccalaureate-results/

Simonite, T. (2020, July 10) *Meet the secret algorithm that keeps students out of college*. Wired Magazine. https://www.wired.com/story/algorithm-set-students-grades-altered-futures/

Wright, E., & Lee, M. (2020). Does the International Baccalaureate ‘work’ as an alternative to mainstream schooling? Perceptions of university students in Hong Kong. *Studies in Higher Education*, *47*(3), 1-16.

1. To maintain anonymity, we just state the initials of each respondent [↑](#footnote-ref-1)