# A consumer perspective of AI certification – the current certification landscape, consumer approval and directions for future research

Received 3 January 2023
Revised 2 June 2023
Accepted 8 September 2023

Myrthe Blösser and Andrea Weihrauch
*Department of Marketing, University of Amsterdam, Amsterdam, The Netherlands*

## Abstract

**Purpose** – In spite of the merits of artificial intelligence (AI) in marketing and social media, harm to consumers has prompted calls for AI auditing/certification. Understanding consumers' approval of AI certification entities is vital for its effectiveness and companies' choice of certification. This study aims to generate important insights into the consumer perspective of AI certifications and stimulate future research.

**Design/methodology/approach** – A literature and status-quo-driven search of the AI certification landscape identifies entities and related concepts. This study empirically explores consumer approval of the most discussed entities in four AI decision domains using an online experiment and outline a research agenda for AI certification in marketing/social media.

**Findings** – Trust in AI certification is complex. The empirical findings show that consumers seem to approve more of non-profit entities than for-profit entities, with the government approving the most.

**Research limitations/implications** – The introduction of AI certification to marketing/social media contributes to work on consumer trust and AI acceptance and structures AI certification research from outside marketing to facilitate future research on AI certification for marketing/social media scholars.

**Practical implications** – For businesses, the authors provide a first insight into consumer preferences for AI-certifying entities, guiding the choice of which entity to use. For policymakers, this work guides their ongoing discussion on "who should certify AI" from a consumer perspective.



*Research involving human participants.* Informed consent was obtained from all individuals who participated in the study.

*Disclosure of potential conflicts of interest.* The authors have no competing interests to declare relevant to this article's content. Financial disclosure. The research leading to these results received funding from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NOW) under Grant Agreement NO. KIVI.2019.006.

*Ethics approval.* The project was approved by the research ethics committee of the University of Amsterdam, Amsterdam Business (Economics and Business Ethics Committee; EC 20220613010611) and performed following the Netherlands Code of Conduct for Scientific Practice. In addition, and per the institution's Research Data Management protocol, all data is available through the institution's data repository.

European Journal of Marketing
Vol. 58 No. 2, 2024
pp. 441-470
Emerald Publishing Limited
0309-0566
DOI 10.1108/EJM-01-2023-0009

**Originality/value** – To the best of the authors' knowledge, this work is the first to introduce the topic of AI certification to the marketing/social media literature, provide a novel guideline to scholars and offer the first set of empirical studies examining consumer approval of AI certifications.

**Keywords** Artificial intelligence, AI certification, Consumer trust, AI regulation, Fair AI

**Paper type** Research paper



Artificial intelligence (AI), defined as "a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan and Haenlein, 2019, p. 15), is rapidly reshaping many domains of life, often assisting and at times even automating human decisions (PwC, 2018). More and more companies are using AI: 50% of companies report AI use in at least one business function (McKinsey, 2021). Fast adoption is often driven by efficiency and economic gains (De Stefano, 2019). AI can be integrated into various business activities and has especially gained importance in the field of marketing. AI can be used to optimise and partly automate marketing channels (e.g. to analyse emails to identify potential leads; Davenport *et al.*, 2020), for personalised dynamic pricing (Hufnagel *et al.*, 2022), or to determine the best time to post social media advertisements (Haleem *et al.*, 2022). AI applications further shape (social) media marketing (e.g. recommendation systems (Shin, 2020) and personalised content in advertisements and streaming services (Hermann, 2021; Valand, 2021)). Lastly, AI is an integral part of marketing-relevant innovations such as voice assistants such as Google and Alexa, smart home applications (Puntoni *et al.*, 2021; Malodia *et al.*, 2022; Philip, 2021), or assistive robots (Bock *et al.*, 2020).

In spite of the above-demonstrated wide-spread use and potential merits, AI has lately gained attention for its potential to harm consumers (Howard and Borenstein, 2018; Mehrabi *et al.*, 2021). Examples also include marketing and (social) media-related AI "failures", such as search engines that show STEM and higher paying job career ads to more men than women (Lambrecht and Tucker, 2019; Datta *et al.*, 2014) and search-engine software that labels photos of black people as primates and black hairstyles as unprofessional for work (Mac, 2021; Prahl and Goh, 2021). It can even lead to unintended price discrimination, such as with Airbnb's smart-pricing tool, which had a disparate impact by ignoring the existing revenue gap between black and white hosts (Zhang *et al.*, 2021).

AI failures can also be found outside of marketing applications, from hiring tools that disadvantage women (Dastin, 2018) to racial biases in criminal justice (Sushina and Sobenin, 2020) or inaccurate health-care diagnoses for black patients (Adamson and Smith, 2018). These AI failures are so prevalent that they are logged on an "AI Incident Database" website by consumers to inform other consumers of their harm (AI Incident Database, 2022).

These incidents are likely to increase because of the growing impact of AI in marketing applications (Vlačić *et al.*, 2021), but also in decision domains such as health (Fan *et al.*, 2020), finance (Cao, 2018), safety (Sushina and Sobenin, 2020) and resource allocation (I Amsterdam, 2020), making consumers increasingly vulnerable. This leads many experts to argue for public policies that ensure Fair, Accountable, Transparent and Explainable (FATE) AI, often referred to as the four pillars of ethical AI (Taeihagh, 2021). As a result, influential organisations such as the Organization for Economic Co-operation and Development (OECD Legal Instruments, 2019) and the European Commission (EC) (EC, 2021) have released regulatory frameworks calling for these principles. The common aim is to guarantee consumers' rights and ensure ethical AI (Jobin *et al.*, 2019).

Nonetheless, critical voices question the success of such frameworks in achieving FATE AI because of a lack of concrete (computational) recommendations and difficulties in

enforcement (AlgorithmWatch, 2021; Veale and Borgesius, 2021). Some suggest explaining the rules behind AI to increase consumer control (Banker and Khetani, 2019). However, understanding AI decisions and underlying (computational) rules requires numerical and analytical ability, which consumers often lack (Castelo *et al.*, 2019; Dietvorst *et al.*, 2015), a phenomenon further referred to as AI illiteracy (Long and Magerko, 2020). This illiteracy renders consumers unable to evaluate AI compliance with FATE principles, which is necessary for contesting discrimination or unfair biases. First educational attempts to improve the AI literacy of non-experts are made; however, these methods are often time-consuming, and their efficacy is yet to be proven (Cheng *et al.*, 2019; Long and Magerko, 2020).

Meanwhile, one of the most actively discussed shorter- and mid-term solutions to warrant FATE AI is AI certification, such as by the US Congress in their algorithm Accountability Act in 2019, The Netherlands Court of Audit in 2020 and research institutions (Mökander *et al.*, 2022; Algemene Rekenkamer, 2022; Guszcza *et al.*, 2018). Certification of an AI, often called *algorithmic auditing*, is "a specific subset of audit studies focused on studying algorithmic systems and content" (Metaxa *et al.*, 2021, p. 6), where the algorithmic system is often tested against a regulatory framework (Cihon *et al.*, 2021).

These auditing strategies are discussed as becoming mandatory for higher-risk applications (i.e. governmental resource allocation, at least in the EU; Lilkov, 2021), but will likely become an optional practice for "lower-risk" for-profit AI applications, giving companies the freedom to choose (Stuurman and Lachaud, 2022). Lower-risk applications include AI used by banks for financial products (i.e. to calculate insurance premiums) but also extend to marketing and social media-related AI use (i.e. personalised recommendations on (social) media platforms, targeted advertisement or pricing). The quest to use AI in these examples to optimise and personalise content for consumers can carry risks. More and more "low-risk" applications are being called out for their potentially harmful practices and inadequate mechanisms to prevent them (Park *et al.*, 2022). An independent audit of Facebook by civil rights attorneys in 2020 states that the platform's efforts to detect algorithmic bias fell short, and consumers stormed to Twitter after discovering that Apple's credit card afforded less credit to women than men (O'Brien, 2020; Vigdor, 2019). As media coverage of an AI failure can lead to reputational harm to the company (Vincent, 2019), certification can not only be instrumental to FATE AI but is also imperative for a company to signal fair AI use to consumers to address calls for social justice (Garcia-Garcia *et al.*, 2021). However, one of the big open questions for AI certification is "Who should certify AI"? The EC, for example, argues for an *independent body* to certify AI, which the European

Economic and Social Committee (EESC) was equally voiced in 2019 (Lilkov, 2021; EESC, 2019). Yet, it is unclear who this independent body should be (Lilkov, 2021) and, more importantly, whom consumers trust to certify AI. Other spaces requiring certification to protect consumers, such as privacy, organic food and fair trade, consist of different players – i.e. governmental institutions, NGOs, commercial third parties or self-declarations to an industry standard – and often get dominated by one entity over time. This process accompanies tensions between for-profit players and public policy involvement until a certification type is established and leaves companies with a transition period of choice (see Web Appendix 1 for an overview of certifying organisations for different sectors). Looking at the real-life landscape of AI certification, we can already observe such tensions emerge with other for-profit and non-profit players entering the field (see pp. 10-12).

Whereas scholars have extensively covered the trustworthiness of certification labels and entities in other fields (Kim and Kim, 2011; Konuk, 2019), previous results from other certification contexts (i.e. organic food and privacy) cannot simply be applied to AI

certification. AI decisions have unique characteristics because of their complexity (Castelo *et al.*, 2019), autonomy (Cunneen *et al.*, 2019) and tremendous scope and decision impact (PwC, 2018; Stone *et al.*, 2016). Nonetheless, research on consumer trust in AI-certifying entities is largely ignored in marketing literature. This is in spite of the reality that trust is crucial for certifications to be effective (Evers *et al.*, 2018) and the documented biases in marketing-related AI applications.

Certification may not completely attenuate biases, but it can help reduce harm to consumers and allow companies to (visibly) demonstrate their commitment to fair AI.

This manuscript, therefore, provides:

- an academic literature and status-quo-driven overview of the AI certification landscape that identifies entities and related concepts (i.e. AI decision domain and type of AI) relevant for consumer approval of AI certification (Part A);
- a first application of the generated concepts in an empirical study answered the question, "Which entities do consumers approve of to certify AI (in different decision domains)?" (Part B); and
- a research agenda for marketing/social media scholars interested in AI trust/fairness and its certification following the identified concepts (Part C).

The above-mentioned research goals serve as a first attempt to guide marketing and social media scholars to better understand what contributes to consumers' approval of AI certifications. Because of its scope, the topic of AI has received attention in many marketing sub-areas (i.e. in social media and marketing channels), often with a focus on *consumer trust* and *AI acceptance* (Longoni *et al.*, 2019; Yalcin *et al.*, 2022). Previous work has examined AI attitudes (Longoni *et al.*, 2019; Yalcin *et al.*, 2022; Kim *et al.*, 2021) or consumer characteristics (e.g. personality traits and AI anxiety; Kaya *et al.*, 2022), consumer innovativeness (Hasan *et al.*, 2021), documented AI harm (Prahl and Goh, 2021; Zhang *et al.*, 2021) and AI applications in marketing (Bock *et al.*, 2020; Vlačić *et al.*, 2021; Davenport *et al.*, 2020). However, AI certification as a possible means to protect consumers and increase consumer trust in AI has not yet been studied (except for a qualitative study examining managers' awareness of AI morality and ethics by Baker-Brunnbauer (2021)). We address this important research gap and focus on one specific dimension more thoroughly: consumers' approval of AI-certifying entities.

In addition to these theoretical contributions, businesses must understand the status quo of the AI certification landscape (Cihon *et al.*, 2021). AI certification helps companies distinguish themselves by informing consumers of their efforts towards FATE AI. Understanding which entities are most approved of by consumers can facilitate the choice of a certifying entity. Ultimately, monetary and time resources are dedicated to a certification process and accompanying (marketing) efforts, and these must align with consumers' preferences. For policymakers, information on consumers' perspectives on AI certification entities can converge opinions and guide the current debate in public policy on who should certify what type of AI (Lima and Cha, 2021).

## Part A: the artificial intelligence certification landscape

In the first step, an academic literature search was conducted. We follow the prospector approach conceptualised by Breslin and Gatrell (2023), which is especially suitable for multi-disciplinary literature searches for intricate real-world problems. Because of the dearth of academic literature on AI certification across disciplines, we chose this over a systematic literature review. We included papers discussing AI certification and possible certification

entities and identified n = 16 relevant manuscripts. The discussed certifying entities were identified and complemented with other relevant factors contributing to consumer approval of and trust in AI certification, such as decision-specificity and type of AI. In a second step, the academic debate was extended by carefully reviewing the status quo of the real-world AI certification landscape.

**Potential certifying entities (academic)**

In line with calls by the US Congress' Algorithm Accountability Act (Mökander *et al.*, 2022), The Netherlands Court of Audit (Algemene Rekenkamer, 2022) and several research institutions (Guszcza *et al.*, 2018), AI certification is an important step in ensuring FATE AI. "Expert" entities ensure the fair use of AI, a role that is currently impossible for consumers because of a generally high level of AI illiteracy (Castelo *et al.*, 2019; Dietvorst *et al.*, 2015). Academic literature identifies several entities to take on this vital task, but in spite of a starting debate, there is no clear consensus (see Table 1). Previous research on AI certification is primarily conceptual; it mainly describes which entities authors favour for the certification process (Erdélyi and Goldsmith, 2018) and recommended procedures for AI certification to follow (Winfield and Jirotka, 2018). The proposed entities are often country or region-specific and cannot be directly translated to other countries that lack a similar infrastructure (Floridi *et al.*, 2018; Roski *et al.*, 2021). To complicate matters, many argue for collaboration between entities, each with different divisions of responsibility (Guihot *et al.*, 2017; Mökander *et al.*, 2022; Floridi *et al.*, 2018; Roski *et al.*, 2021).

In spite of a lack of clear consensus, the academic debate suggests an influential role for the *government*. This is reflected in work of Scherer (2015), Erdélyi and Goldsmith (2018), Floridi *et al.* (2018), Falco *et al.* (2021) and Badran (2021), who argue for certification by the government. Others argue for a collaborative role, urging the government to certify AI and join forces with industry players for standard-setting to ensure a fit with the real-life challenges of AI applications (Tutt, 2017), or the reverse, where the government defines FATE AI standards, but the company itself is responsible for testing their adherence to these standards in the form of a *self-declaration* (Guihot *et al.*, 2017) because of governments' limited resources, or by *NGOs* and *commercial third parties* (Mökander *et al.*, 2022) as part of a broader ecosystem. Not limited to the government, Stuurman and Lachaud (2022) propose a label like the nutritional label in the EU for *NGOs* and *commercial third parties*.

Others propose solely for-profit entities, such as a *commercial third party*, to certify AI (Yanisky-Ravid and Hallisey, 2019; Arnold *et al.*, 2019; Raji *et al.*, 2022). Commercial third parties would allow for independent oversight without directly profiting from the AI they certify (Raji *et al.*, 2022). In addition, auditing is often the core business of a commercial third party and therefore has existing infrastructure and expertise. Alternatively, some propose *independent bodies* as an all-encompassing term, including *NGOs* and *governmental* agencies such as the FDA. Much like Raji *et al.* (2022), they argue that an independent body ensures the quality of the certification (Sharkov *et al.*, 2021) and the competence of the auditors (Winter *et al.*, 2021).

This independence is optional for Winfield and Jirotka (2018) and Roski *et al.* (2021), who propose *self-declarations*. Companies themselves would be responsible for ensuring their efforts towards FATE principles without external oversight, arguing that external bodies often don't have enough expertise (i.e. the government) and cannot keep up with the pace of AI innovation (Roski *et al.*, 2021).

**Table 1.**
Overview of literature outside of marketing on AI certification

| Authors | Year | Field | Methodology | Type/part of AI | Entity proposed | Geographical focus | Mention of consumer trust | Decision specificity | | Summary |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Domain | High-risk/ low risk | |
| Scherer | 2015 | Law | Prescriptive | Automated machines | Government | USA | | | | Discusses the complexity of governmental regulation of AI and proposes a framework to deal with these challenges |
| Guihot *et al.* | 2017 | Law | Prescriptive | AI | Government and self-declaration | Global | | ✓ | ✓ | Urges for the government's involvement in standard setting for AI applications and self-declarations by industry players |
| Tutt | 2017 | Law | Prescriptive | AI | Government | USA | ✓ | | | Argues for an FDA-like institution to certify AI applications and develop AI standards in partnership with the industry |
| Erdélyi and Goldsmith | 2018 | Computer Science | Prescriptive | AI | Government | Global | | | | Urges for a new international governmental body for standard-setting that can inform individual countries' regulations and legislation |
| Winfield and Jirotka | 2018 | Computer Science | Prescriptive | Intelligent Autonomous Systems | Self-declaration | Global | ✓ | | | Outlines a roadmap to ethical governance, including a company's self-declaration of ethical conduct |

(*continued*)

Table 1.

| Authors | Year | Field | Methodology | Type/part of AI | Entity proposed | Geographical focus | Mention of consumer trust | Decision specificity: Domain | Decision specificity: High-risk/low risk | Summary |
|---|---|---|---|---|---|---|---|---|---|---|
| Floridi *et al.* | 2018 | Philosophy and Cognitive Sciences | Prescriptive | AI | Government | European Union | ✓ | | | Provides concrete recommendations for AI governance and certification within the European Union |
| Yanisky and Hallisey | 2019 | Law | Prescriptive | Data used for the development of AI | Government or (commercial) third-party | USA | | ✓ | | Proposes an "AI Data Transparency Model". The model includes an auditing regime and certification program |
| Arnold *et al.* | 2019 | Industry Paper | Prescriptive | AI services for developers | Self-declaration and (commercial) third-party | Global | ✓ | | | Proposes and outlines the characteristics of a supplier certificate called FactSheet |
| Sharkov *et al.* | 2021 | Information and Security+ | Descriptive | AI | Independent body | European Union | ✓ | | | Provides an overview of European initiatives towards AI regulation/certification and argues for an independent body to certify AI |
| Winter *et al.* | 2021 | Computer Science | Prescriptive | Low-risk machine learning applications | Independent body | Global | | | ✓ | Outline existing standardisation methods used in other fields and introduce four levels relating to AI's impact on people, the environment and organisations to be certified |

(*continued*)

Table 1.

| Authors | Year | Field | Methodology | Type/part of AI | Entity proposed | Geographical focus | Mention of consumer trust | Decision specificity | | Summary |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Domain | High-risk/ low risk | |
| Falco *et al.* | 2021 | Computer Science | Prescriptive | Automated systems | Government | Global | ✓ | | | Argue for certification of automated AI systems according to three "AAA" governance principles: Assessment of risk; Audit trails; and system Adherence to jurisdictional requirements |
| Badran | 2021 | Computer Science | Descriptive | AI | (Governmental) independent body | Global | | | | Provides an overview of the pitfalls of governmental AI regulation and opportunities for an independent regulatory agency to certify AI |
| Roski *et al.* | 2021 | Medicine | Prescriptive | AI | Self-declaration | USA | ✓ | ✓ | | Outlines a process for the health-care industry to develop standards for AI certification to promote more trust |
| Mökander *et al.* | 2022 | Philosophy and Cognitive Sciences | Descriptive | High-risk AI | Government and NGO or (commercial) third-party | European Union | | | ✓ | Provides an overview of the proposed European Artificial Intelligence Act (AIA) and specifies how the proposed conformity assessments could be applied using existing literature on AI auditing |

(*continued*)

| Authors | Year | Field | Methodology | Type/part of AI | Entity proposed | Geographical focus | Decision specificity | | | Summary |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Mention of consumer trust | Domain | High-risk/ low risk | |
| Raji *et al.* | 2022 | Computer Science and Law | Descriptive | AI | (Commercial) third-party | USA | | | | Argues that the current landscape does not allow for effective commercial third-party algorithmic auditing and argues for solutions by drawing a connection to other certification spaces (e.g. finance, health) |
| Stuurman and Lauchaud | 2022 | Law | Prescriptive | AI | NGO or (commercial) third-party | European Union | ✓ | | ✓ | Outlines the requirements for a voluntary label for medium- to low-risk AI applications and the challenges in the implementation |

**Source:** Authors' own work

**Table 1.**

### Potential certifying entities (real world)

To complement the entities mentioned in the academic debate while acknowledging the dynamic nature of the landscape in AI certification and similar areas, we also look at the actual state of the AI certification market. Several "players" are entering the field, including public policy initiatives and for-profit organisations.

*Governmental bodies* are taking first steps towards the regulation and certification of AI, such as the cities of Amsterdam and Helsinki, which established an AI registry that can be used to evaluate and certify different AI decisions (AI for Good, 2023). Another example is the Danish Ministry of Finance, which launched an IT certification program that checks responsible data use (Danish Ministry of Industry, Business, and Financial Affairs, 2019).

*Commercial third parties (*e.g. *auditing companies)*, such as ORCAA, already operate in AI auditing. ORCAA was founded by Cathy O'Neil, a researcher famous for her book "Weapons of Math Destruction" (O'Neil, 2017) and a loud voice for regulating AI. ORCAA offered the first-of-its-kind certification to HireVue in 2020 (Zuloaga, 2021). While not yet offering certification, auditing companies Deloitte and KPMG provide consulting for fair and transparent use of AI systems within companies using their Trustworthy AI framework (Deloitte, 2023) and AI in Control framework (KPMG, 2023).

While these organisations offer services to a wide range of algorithmic applications within the public and private sectors, other initiatives from *not-for-profit entities and NGOs* are narrower in scope, such as FairAI, a volunteer organisation within Oxford University. FairAI is developing a method using blockchain technology to certify and verify companies' efforts in mitigating workplace displacement by algorithms (2023). Another example is the Certify-AI service offered by Columbia University's Data Science Institute, which tests whether an algorithm meets the standards issued by the FDA for health-care applications. Focusing on AI applications within telecommunications and electrical engineering, many institutions – mostly *not-for-profit* – are standardising the development and deployment of AI, both on a national and international level, as mapped in detail by de Winter *et al.* (2021) and Cihon *et al.* (2021). While these initiatives are a critical step in the right direction, they are scattered in focus, often non-generalisable to other fields and, more importantly, developed and offered by different entities.

Lastly, and probably most concerning, current AI decisions *lack a certification* altogether or are limited to a *self-declaration* to the company-launched standards, such as done by Google (2023), IBM (2019) and Microsoft (2018), all of whom developed and published their own AI principles (Jobin *et al.*, 2019). However, De Laat (2021) found that many self-declaring companies lack concrete actions, making many declarations untrustworthy (Stuurman and Lachaud, 2022).

Besides these entities, there are other initiatives aimed at ensuring quality assurance, such as Watchdogs (Ada Lovelace Institute, AI Now, and Open Government Partnership, 2021), ethical committees within technology companies (Candelon *et al.*, 2022), decentralised data access (Baird and Schuller, 2020), algorithmic social contracts (Rahwan, 2018) and human rights impact assessments (Moss *et al.*, 2021). However, these initiatives are less known to the public and do not communicate the company's efforts towards FATE AI to the consumer.

### Other important factors

In addition to proposing entities, academic literature discusses factors related to the question of who should certify AI. Among them is the decision-specificity of the AI application. This dimension considers whether an application is "high-risk" or "low-risk" and the domain in which the AI is used.

*Decision-specificity – high-/low-risk application.* The possibility for AI applications is immense and spans many domains, ranging in the risk posed to the consumer. The EC's Draft AI recognises this by imposing different regulations for high- and low-risk AI applications, with higher risk applications facing stricter regulation (Lilkov, 2021). High-risk AI includes the impact on the consumer, i.e. health risks, safety or even human rights, such as its use in justice systems and health diagnosis (Sushina and Sobenin, 2020; Fan *et al.*, 2020), but could also apply when citizen fairness is at stake – i.e. in resource allocation for social housing or when used by tax authorities. Consumers' inability to opt out reflects this high-risk such as when a governmental institution uses it. In contrast, low-risk applications are often associated with a lower impact on the consumer. This is partly because of the freedom of choice; consumers actively decide to engage with certain companies, products or social media platforms. As such, marketing AI applications are often low-risk.

Consumers might reflect the level of risk in their approval of different entities in different domains. Governmental regulatory agencies are involved in other high-risk applications, such as those that pose a risk to life (e.g. the FDA regulates the safety and effectiveness of drugs and medical devices in the USA). However, there may be conflicts of interest when a government agency regulates its actions, like in public safety (e.g. policing). In these cases, consumers may prefer an independent organisation such as NGOs or commercial third parties.

*Decision-specificity – decision domain.* Additionally, different domains (i.e. sectors such as finance or public health) require different methods, entities and possibly standards for AI certification (Raji *et al.*, 2022). The process of certifying AI applications in health care, such as diagnosing an abnormality in an MRI with AI, vastly differs from that of applications in the financial sector, such as calculating mortgages. As a result, consumer preferences and trust in certification entities may not transfer between domains, especially between the public sphere, where the consumer cannot opt out (e.g. the government), and private companies, where consumers have a choice.

Besides, consumer trust is a multifaceted concept that is influenced by a variety of factors, including the credibility and expertise of the entity (Lanero *et al.*, 2021; Janssen and Hamm, 2012; Roe and Teisl, 2007), as well as the consumer's familiarity with the entity (Brach *et al.*, 2018). As the AI landscape is yet to be defined, consumers may rely more on their perceived expertise than their familiarity, especially as AI requires specialised expertise to understand. Domination of specific entities for other applications within those domains might signal expertise to the consumer, such as commercial third parties' domination in the financial sector, the government in the health and safety domains and governmental agencies and NGOs for social domains (i.e. resource allocation) (Nawaz, 2018; UK Government, 2023).

## Consumer approval/trust

Examining the literature and the status quo provides insights into potential entities, but preferences may depend on decision risk and domain. What is largely absent in existing research endeavours is the perspective of those who ultimately must approve of the certifications: the consumer. Law and computer science scholars dominate the literature, and while some acknowledge the importance of consumer trust, most adopt a top-down approach that ignores further analysis of the consumers' opinions. Marketing literature is lacking altogether. A survey by Ipsos found that only half of participants from 18 countries believe AI has more benefits than drawbacks, suggesting consumers may seek assurance (Ipsos, 2022). Not only will certification protect the consumer from potential harm, but it also allows for greater trust in AI in general (Knowles and Richards, 2021).

This lack of consumer perspective is even more problematic as other fields, such as organic food and privacy, have rich literature that recognises that consumer trust is important and crucial for certification effectiveness (Evers *et al.*, 2018; Kim and Kim, 2011; Konuk, 2019). Meanwhile, AI is expanding into more marketing-related functions such as information filters on social media, pricing algorithms, search engine recommendations and advertisements that not only have shown the potential to cause harm (Arora *et al.*, 2022; Pandey and Caliskan, 2021; Kay *et al.*, 2015; Datta *et al.*, 2014) but are also a day-to-day reality for many consumers. Evidently, there is yet to be a clear answer to who should certify AI from a consumer perspective.

**Part B: research choices deducted from the table**

The second part of this manuscript serves two purposes: it provides an example of how to apply the concepts deducted in Part A to empirical studies from a consumer perspective and a first understanding of which entity consumers may approve of to certify AI applications in multiple domains (social, safety, health and finance). By doing that, we address one of the most important gaps in the current academic work. As shown in Column 4 of Table 1 ("methodological approach"), most scholars have *descriptively* outlined the processes for AI certification. Others are *prescriptive* and provide specific roadmaps and schemes (Winfield and Jirotka, 2018; Arnold *et al.*, 2019; Falco *et al.*, 2021). However, empirical work is absent. As a result, we opted to specifically cover the proposed entities (Column 6), decision specificity (Column 9), namely, different decision domains with varying levels of risks and initiating parties and consumer trust (Column 10). Our empirical approach was therefore strongly guided by using identified dimensions/columns of Part A. Below, we outline the included dimension columns in more detail before providing methodological details on our study.

*Entity proposed (Column 6).* To account for the entities proposed in academic literature and the real-world first players identified in Part A, we include a governmental institution, NGO, commercial third party and self-declaration. These players allow us to understand consumers' potential associations with these entities and cover the current AI landscape well.

*Decision specificity (Column 8).* The popularity and diffusion of AI led to AI being used in multiple domains and for decisions with very different risk profiles. Nonetheless, most of the literature in Part A does not account for potential differences between these decisions regarding AI certifications and the entities involved (also see pp. 12-14). While it is impossible to cover the entire complexity of the vast range of AI decisions, we chose domains and specific AI decisions based on reports examining in which areas AI is currently most dominantly used and projected to have a considerable influence in the future for consumers (PwC, 2018; Stone *et al.*, 2016). As a result, we include four application domains: social, safety, health and finance. These domains include low- and high-risk applications, allowing us to cover both AI applications within the public sphere (e.g. social housing allocation) and the private sphere (e.g. the height of insurance). In spite of not explicitly including marketing/social media AI application domains, we believe that varying decision specificity allows us to gain insights for marketing while maximising the impact based on projected spread and influence.

*Consumer trust (Column 10).* While some of the papers identified in Part A mention the importance of consumer approval/trust, they have yet to examine this important variable empirically. There are two empirical, noteworthy exceptions outside AI certification, looking at consumers' perceptions of entities *developing and governing AI*, which show tensions between for-profit and not-for-profit entities. Zhang and Dafoe's (2019) survey with a

representative sample of 2000 participants showed that university researchers and the US military are the most trusted entities to develop AI, followed by tech companies and NGOs, with the government being the least trusted. However, the authors also report a general distrust of any institution using or developing AI among the participants.

KPMG and the University of Queensland partly corroborate these findings in their survey with 6,054 participants from the USA, Canada, Germany, the UK and Australia (Gillespie *et al.*, 2021), showing that universities and other research bodies are most trusted to develop (and govern) AI while accounting for differences between countries, and governments and commercial third parties are trusted the least. While these two surveys offer the first insights into consumers' opinions, the entities provided in the surveys do not represent the current parties already entering the field or those proposed by academic literature. More importantly, they look at AI development and not AI certification. As different skill sets, motives and consumer associations are linked to the development process versus audit/certification of AI (Raji *et al.*, 2022), the results are not directly applicable to the certification landscape.

## Empirical design

*Pilot.* We first ran a pilot study with a five-entity type (governmental institution, NGO, commercial third-party, self-declaration and no certification) × four-domain (social, safety, health and finance) design, with both entity type and domain as within-subject factors. In total, 302 participants served as a sample (70.3% female, $M_{age} = 36.73$ years, SD = 13.91). The study procedure, questionnaire and results of the pilot can be found in Web Appendices 2, 3 and 4.

We made several changes to the main study based on the pilot study's results and participant comments. We:

- used a representative sample to ensure a balanced sample;
- excluded the option for "no" certification efforts as this option generally had low approval;
- removed specific examples of entities (i.e. KPMG, a large auditing firm, as an example for a commercial third party) as we acknowledge that participants may hold biases or associations with these examples; and
- changed the entity factor to a between-subjects design to prevent scenario fatigue and mitigate the potential drawbacks of a within-subject design.

*Main study.* As a result, the main study adopts a four-entity type (governmental institution, NGO, commercial third-party and self-declaration to industry standard) between-subject design with four-domain (social, safety, health and finance) within-subject factor and addresses the shortcomings of the pilot study (also see procedure).

## Participants

All the participants were recruited via Prolific (www.prolific.co/). Participants received a monetary reward in line with the platform's requirements at the time of data collection (£5 per hour). We opted for a representative sample (gender, age and ethnicity) and only accepted participants with an approval rate of over 90%. We aimed for n = 100 for each between-subject cell, resulting in 404 participants (51% female, $M_{age} = 44.59$ years, SD = 15.6). To determine our sample's geographical region, we first identified the region for which AI certification is most advanced. Considering progress and clusters of guidelines (Jobin *et al.*, 2019), Europe was identified. Within Europe, we decided to test defensively. Based on

the pilot study results, which indicated preferences for non-profit entities, specifically governmental institutions, we identified the UK as the country with the lowest consumer trust in the government (Davies *et al.*, 2021; OECD Data, 2022). In addition, the UK has the highest AI readiness index score in Europe and populates second place globally (Oxford Insights, 2019), likely accelerating AI diffusion and the need for certification. We report all data exclusions and measures, and all data sets are available upon request.

## Procedure

After giving consent, all participants viewed a short clip (1,5 minutes) explaining the basic principles of AI to ensure a baseline understanding. To establish a uniform perception of the certification process for every entity involved, participants then read a text saying that the certification entity would test the AI decision against pre-defined international standards by the Organization for Economic Cooperation and Development to ensure FATE AI. They then read a short text outlining these rules (1 minute; OECD Legal Instruments, 2019) (see Web Appendix 5 for the full scenario). Afterwards, participants were exposed to eight AI decisions in random order from four domains (see Table 2 for an overview of all decisions). We chose two sample decisions for each domain for generalisability. A composite score was created for each domain by averaging the respective two AI decisions. The individual AI decisions resembled those of their respective domains.

To ensure that the AI decisions were relevant to consumers, we included two measures assessing the impact these decisions would have on their life (1 = not at all impactful and 7 = very impactful) and the likelihood that these decisions would affect their own life (1 = extremely unlikely and 7 = extremely likely) (see Web Appendix 6 for the full questionnaire and Web Appendix 7 for statistics).

For each of the AI decisions, participants were then asked to judge their approval for one entity (e.g. "To what degree would you approve if you were informed that the [*government]* was the entity who was responsible for the certification process of using artificial

| Domain | Decision scenario | Explanation |
|---|---|---|
| Social | Social housing allocation | The use of artificial intelligence to decide who should get social housing (often from a limited number of houses available) |
| | Pension allocation | The distribution of pensions across all retirees |
| Safety | Real-time analysis of safety cameras | Artificial intelligence program that analyses videos from safety cameras and predicts the need for emergency vehicles |
| | Predictive crime policing | Artificial intelligence decides the police headcount in a certain area, leading to more policing in certain areas than others |
| Health | Medical imaging diagnostics | The use of artificial intelligence to detect abnormalities in radiology images |
| | Public health prediction | Artificial intelligence that predicts future public health events (such as a flu outbreaks) |
| Finance | Height of mortgage | Risk assessment that leads to higher mortgage prices for individuals with higher risk assessments |
| | Price of insurance (i.e. car insurance or liability insurance) | Risk assessment that leads to higher insurance prices for individuals with higher risk assessment |

**Source:** Authors' own work

**Table 2.**
AI decisions provided to the participants with explanations

intelligence for the examples below?") on a seven-point Likert scale, where 1 = very much disapprove and 7 = very much approve, resulting in an eight-item scale. All decisions were presented in a random order. This led to the following allocation of participants: governmental condition ($N$ = 100), commercial third-party condition ($N$ = 100), NGO condition ($N$ = 102) and self-declaration condition ($N$ = 102).

Participants were then asked to report: age (in years); gender (female, male or other); the highest level of obtained education (elementary school, high school, bachelor's, master's, PhD); political orientation (sliding scale from 0 to 100, where we provided labels at 0 (the Conservative Party), 50 (Liberal Democrats) and 100 (Labour Party); and participants' own AI expertise (on a seven-point Likert-scale with a three-item questionnaire such as "I feel that I know more about artificial intelligence than others", adopted and fit to the AI-context by Park *et al.* (1994), where 1 = strongly disagree and 7 = strongly agree). We included these control variables as previous research shows that they impact attitudes towards AI (e.g. older and more left-wing citizens show stronger support for AI regulation [O'Shaughnessy *et al.*, 2021; Wang *et al.*, 2020; Zhang and Dafoe, 2019]), as well as the assessment of the included entities (e.g. women support the government more than men, while men trust for-profit companies more than women [Christensen and Lægreid, 2005; Pirson *et al.*, 2019]).

Additionally, political orientation can lead to differences in approval for AI use in different domains. For example, conservative citizens are found to be more approving of predictive policing than more liberal citizens (O'Shaughnessy *et al.*, 2021). Lastly, participants were given the option to leave any comments regarding the AI decisions, the entities, or the future of AI in general in a textbox. We included this textbox in the study because the literature review in Part A identified non-entity-related technological/computational quality assurance processes. While participants did not mention these in the open-ended question in the pilot study, we wanted to determine if our representative sample had a similar knowledge base of current initiatives in AI regulation.

## Results and discussion

In a first attempt to guide our research choices using the literature overview, we structured our analyses according to the different dimensions.

*Different entities (Column 6).* A one-way ANOVA was conducted to test for possible differences in the approval of the entities. To correct for the assumption of homogeneity not being met [$F(3, 400) = 6.252$, $p < 0.001$], results were read from Welch's robust test of equality of means, which yielded significant differences in the approval for the different entities, $F(3, 222.35 = 4.063$, $p = 0.008$, $\omega^2 = 0.058$. Tukey post hoc tests revealed that approval for the government ($M_{\text{government}} = 4.79$, SD = 1.25) is significantly higher than that of the commercial third party ($M_{\text{commercialthirdparty}} = 4.19$, SD = 1.72, $p = 0.033$) and the self-declaration ($M_{\text{selfdeclaration}} = 4.17$, SD = 1.76, $p = 0.023$) (see Figure 1). It is, however, not significantly different from those for the NGO ($M_{\text{NGO}} = 4.47$, SD = 1.42, $p = 0.461$). There is no significant difference in approval between the other entities.

The results of this study slightly differ from the within-subject pilot study, in which there was a more pronounced difference in approval between the different entities and a clear(er) tendency towards non-profit (government and NGO) versus for-profit (commercial third-party and self-declaration) entities. We, therefore, further collapsed the entities into approval for non-profit and for-profit entities to understand whether the current study shows a similar trend. Results revealed that approval is significantly higher for non-profit entities ($M_{\text{non-profit}} = 4.63$, SD = 1.35) than for-profit entities ($M_{\text{for-profit}} = 4.18$, SD = 1.74), $t(402) = 2.924$, $p = 0.004$, in line with the findings from the pilot study.

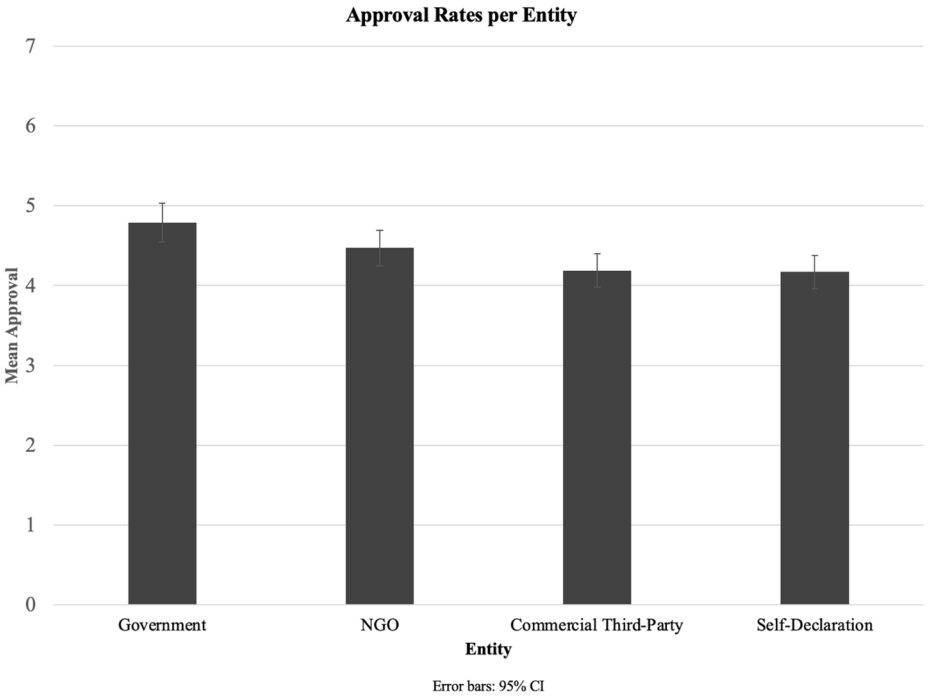
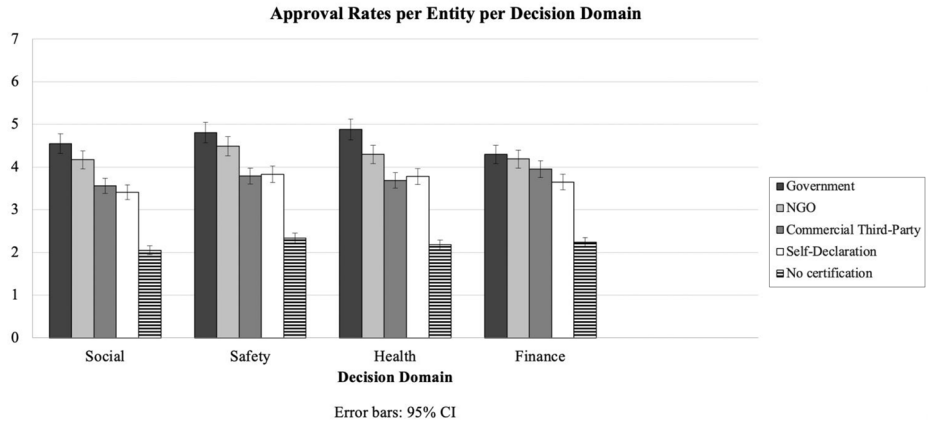



**Figure 1.**
Mean levels of approval for each entity

**Source:** Authors' own work

*Domain specificity (Column 9)*. We carried out a mixed ANOVA with different entities as a between-subject factor (*governmental institution*, *commercial third-party*, *NGO* and *self-declaration*) and the AI decision domains as a within-subject factor (*social*, *safety*, *health* and *finance*; see Figure 2) to better understand whether domain-specificity influences consumer approval for different entities. All approval rates per AI decision and entity can be found in Appendix. Because of the violation of the assumption of sphericity, as indicated by Mauchly's test, $\chi^2(5) = 0.89$, $p < 0.001$, and the Greenhouse–Geisser value being higher than 0.75, we applied a Huynh–Feldt correction. Results show a significant main effect of the domain [$F(2.82, 1127.65] = 26.08$, $p < 0.001$, $\eta_p^2 = 0.06$ as well as a significant interaction effect between the domain and the different entities, $F(8.46, 1127.65) = 2.45$, $p = 0.011$, $\eta_p^2 = 0.018$. Bonferroni-corrected pairwise comparisons show that the approval of the different entities did not differ for the social and financial domains. However, approval of the government as a certifying entity in the health domain is significantly different from that for the commercial third-party ($p = 0.019$) and the self-declaration ($p = .018$) but is not significantly different from that for the NGO ($p = 1.000$). The same pattern can be seen for the safety domain, where approval for the government is significantly higher than that for the commercial third party ($p = 0.001$) and the self-declaration ($p = 0.001$). It is, however, not significantly higher than the approval for the NGO ($p = 0.275$). All other entities in the health and safety domain do not significantly differ from each other. All pairwise comparisons can be found in Web Appendix 8.




**Source:** Authors' own work

**Figure 2.** Mean levels of approval for each entity for each domain

We repeated analyses from the *entity* and *domain* columns with the control variables: age, gender, education (1.2% elementary school, 33.7% high school, 43.3% bachelor's, 13.9% master's, 2.2% PhD, 5.7% other), political orientation ($M = 57.40$, SD = 28.75) and their own AI expertise ($M = 3.27$, SD = 1.31). Only AI expertise significantly affected consumer approval but did not alter the approval for the entities or the different domains ($p = 0.002$).

*Discussion.* The results show that consumers' approval differs between the entities, with an overall preference for the government as a certifying entity. To an extent, consumers also seem to distinguish between public (non-profit; government and NGOs) entities and private (for-profit: commercial third parties and self-declaration to industry standards) entities. This is in line with Gillespie *et al.* (2021), where commercial third parties were viewed as developing and regulating AI for financial gains – that is, consumers held anti-profit beliefs (Bhattacharjee *et al.*, 2017). Surprisingly, approval for self-declarations was not significantly different from approval for commercial third parties, in spite of the lack of payment or external oversight for the former. This may be because of consumers' limited understanding of certification processes, similar to organic food certification (Grunert *et al.*, 2014). Although we aimed to establish a baseline understanding of certification in our scenario, participants may have still struggled to differentiate between third-party and self-declaration processes. For example, participants may believe that a self-declaration signifies that a company has internal controls for compliance (e.g. an ethical board) and is, therefore, similar to the processes followed by a commercial third party. In our General Discussion section, we further discuss this and other alternative explanations and study limitations in more detail.

In addition, decision specificity is an important factor for consumers' approval. Consumers prefer the government over for-profit entities in the health and safety domain but not in the social and financial domains. One explanation could be that governmental agencies mostly regulate the health and safety domain in the UK (Nawaz, 2018; UK Government, 2023), possibly signalling existing expertise and credibility in these domains (Lanero *et al.*, 2021; Janssen and Hamm, 2012; Roe and Teisl, 2007). The social and finance domains are coregulated in the UK, possibly leading to the participants' lack of clear preference for these domains. However, results do show that consumers recognise that different domains require different methods, entities and standards for their AI certification (Raji *et al.*, 2022).

## Part C: future research agenda for marketing/social-media scholars

The previous part served as a first empirical exploration of how to deduct research gaps in AI certification based on the academic overview and the AI landscape's status quo. The pilot study and the main study show that consumers prefer different entities for the certification of AI and these preferences differ for different domains. This last part outlines additional ways our table can guide and inspire (marketing) research. While not encompassing all possible avenues, the below-mentioned areas serve as a starting point (see Table 3 for exemplary related research questions for each column).

### Type/part of artificial intelligence (Column 5)

AI consists of different types (i.e. rule-based AI or machine learning), and AI systems that lead to an AI decision could be composed of different parts (i.e. trained data and AI code).

The complexity of these AI types and high AI illiteracy in understanding these complexities (especially considering upcoming laws; Cheng *et al.*, 2019) open important research avenues for marketing scholars. Line 1 of Table 3 provides exemplary questions on AI certification research focused on AI types/parts.

### Entity proposed (Column 6)

This paper maintains a consumer perspective and, therefore, needs to take a level of abstraction that aligns with the consumers' knowledge of the AI certification landscape. However, to cover the complexity of the landscape and to better assist companies' communication of their certification efforts, a deeper investigation into a wider granularity of entities, the inclusion of other AI quality assurance strategies, consumers' pre-existing attitudes towards the entities and consumer characteristics (i.e. political orientation) would be instrumental (Stuurman and Lachaud, 2022; Cihon *et al.*, 2021). Line 2 of Table 3 presents exemplary research inquiries on the proposed AI certification entities.

### Decision specificity (Column 9)

Although we prioritised domains with the most impact (PwC, 2018; Stone *et al.*, 2016), consumers may encounter AI more frequently in other domains, such as social media. However, consumers often underestimate the prevalence of AI applications in their daily lives, as reflected by naïve consumers having less nuanced attitudes towards AI governance between domains compared to AI experts (Zhang and Dafoe, 2019; O'Shaughnessy *et al.*, 2021). Future research efforts into educating consumers about AI's harms and applications in daily life can provide more nuanced insights on consumer approval for AI certification in different domains. Line 3 in Table 3 outlines further research questions related to the decision-specificity of AI decisions.

### Consumer trust (Column 8)

To account for situations where consumers cannot opt out (e.g. AI decisions by the government), we focused on consumer approval as a measure of trust. However, a more detailed measurement of trust accounting for its various dimensions could provide insights into "why" consumers approve of one entity over another (Roe and Teisl, 2007; Janssen and Hamm, 2012). This, together with a deeper investigation into consumer characteristics and more downstream consequences (Wang *et al.*, 2020; Vincent, 2019), could provide more insights into the construal of consumer trust in AI certification. Line 4 in Table 3 posits different research inquiries on AI research focused on consumer trust.

| Column | Research questions |
|---|---|
| Type/part of AI (Column 5) | • Are there differences in consumers' certification preferences depending on. . .<br>– . . . type of AI (i.e. rule-based, machine learning and neurological networks)?<br>• Can certification be replaced/perceived as less relevant if companies apply. . .<br>– . . . debiasing data strategies, computational fairness metrics, human governance?<br>• Are consumers able to assess certification regarding these AI-related specificities?<br>– How can certification efforts be communicated in conjunction with the complexity of AI types/parts? |
| Entity (Column 6) | • Does consumer approval differ between. . .<br>– . . . more granular options within the entities we proposed (i.e. defence forces, intergovernmental research institutions, university research bodies and technology companies)?<br>– . . . different AI quality assurance strategies outside of certification (i.e. decentralised data access, algorithmic social contracts (Baird and Schuller, 2020; Rahwan, 2018))?<br>• What is the role of pre-existing attitudes towards entities in consumers' approval for AI certification? (i.e. trust and perceived morality).<br>• How do consumer characteristics, such as socio-demographics (e.g. age and gender), political orientation and level of education affect their approval of different entities)? |
| Decision specificity (Column 9) | • How can you successfully educate consumers on the prevalence of AI in their daily life? (i.e. in domains such as social media and recruitment)<br>– How does this, in turn, influence their attitudes towards AI certification and the certifying entities?<br>• Which factors lead consumers to differ in their approval for different domains?<br>– i.e. levels of risk, perceived impact and interaction frequency? |
| Consumer trust (Column 8) | • How do the different dimensions of trust influence consumer approval?<br>• What are the downstream consequences of different AI certifications for companies (e.g. purchase intention, word-of-mouth)?<br>– Can negative views of the certifying entity "spill over" to the company getting certified?<br>– If so, how can this spill-over effect be mitigated?<br>• Are there any consumer characteristics that influence consumer approval for the different entities?<br>– i.e. risk aversion, techno-scepticism, general disposition to trust and consumer innovativeness (O'Shaughnessy *et al.*, 2021; Kim and Kim, 2011; Konuk, 2019) |

(*continued*)

**Table 3.** Summary of questions for future research

| Column | Research questions |
|---|---|
| Geographical focus (Column 7) | • Are there differences in consumer approval between individualistic and collectivistic cultures?<br>– i.e. the influence of privacy concern, familiarity with AI and general comfortableness with AI (Wang *et al.*, 2020; Belanche *et al.*, 2019; Wright *et al.*, 2021)<br>• What is the effect of the political structure on the approval for different entities?<br>– e.g. consumers' institutional trust in democratic countries often hinges on the institutions' performance, while consumers from more authoritarian countries such as China are influenced mainly by state propaganda and cultural values (Yang and Tang, 2010)<br>• Are there differences in approval between different cultures and geographical regions? What are the main factors that drive these differences? |

**Table 3.** **Source:** Authors' own work

### Geographical focus (Column 7)

Our studies took a strictly Eurocentric view by only including UK participants. We argue that our approach is defensive; the results indicate the most approval for the government in spite of the lowest institutional trust in Europe (Clery *et al.*, 2021; Davies *et al.*, 2021). However, consumers' construction of trust and attitudes towards AI differ greatly between countries (Yang and Tang, 2010). Not only will future research benefit from the inclusion of countries with different political structures and cultural values as it adds to the generalisability of the results, but it also addresses calls for inclusive and participatory AI governance (O'Shaughnessy *et al.*, 2021) and provides marketers with the right tools to tailor their certification efforts per region and country. Line 5 in Table 3 presents research questions focused on the geographical focus of AI research.

## General discussion

Recently, AI applications have received considerable attention for their potential to harm consumers and display bias against certain consumer groups (Park *et al.*, 2022; Howard and Borenstein, 2018; Mehrabi *et al.*, 2021). Examples of marketing AI "failures" include search engines showing STEM and higher paying job career ads to more men than women (Lambrecht and Tucker, 2019; Datta *et al.*, 2014) or (unintended) price discrimination against black Airbnb hosts (Zhang *et al.*, 2021). In social media marketing, AI's self-reinforcing nature has received attention for creating echo chambers, ultimately leading to polarisation (Cinelli *et al.*, 2021; Arora *et al.*, 2022). AI failures can also be found outside of marketing applications, from hiring tools that disadvantage women (Dastin, 2018) to racial biases in criminal justice (Sushina and Sobenin, 2020) or inaccurate health-care diagnoses for black patients (Adamson and Smith, 2018).

Examples like these have led scholars and experts to push for AI certification as one solution to protect consumers (Guszcza *et al.*, 2018). The push for AI certification, defined as "a specific subset of audit studies focused on studying algorithmic systems and content" (Metaxa *et al.*, 2021, p. 6), where the algorithmic system is often tested against a regulatory framework (Cihon *et al.*, 2021), is emphasised by the general low AI literacy of consumers (Long and Magerko, 2020). Some name AI certifications as one of the most important strategies to protect consumers while ensuring consumer trust in AI (and its merits). However, AI certification has been mostly overlooked by marketing scholars. As any

certification is only as good as the consumers' trust in it, the dearth of marketing literature on AI certification is an oversight, especially as marketing scholars are specifically suited to lead efforts to better understand consumer trust in and approval of AI certifications (Shin, 2020).

This manuscript, therefore, introduces the topic of AI certification to the marketing/social media literature. By reviewing work outside of marketing and examining the status quo of the AI certification landscape, we identify entities and related concepts (i.e. methodology, type/part of AI, geographical focus and decision-specificity) relevant to consumer approval of AI certification (Part A).

We then provide a first empirical application of the generated concepts by answering the question, "Which entities do consumers approve of to certify AI (in different decision domains)?" (Part B). Results of a pilot (within-subject design) and the main study (between-subject design) reveal differences in consumer approval for four certifying entities (government, NGO, commercial third-party and self-declaration) in different decision domains (social, safety, health and finance). Results show that consumers call for governmental institutions (and NGOs) to claim their space. The exploratory study also points to interesting differences between domains (especially in health and safety). As businesses are currently free to select their certification entities, these results point to governmental institutions or NGOs as the best choice. This insight is equally important for policymakers as several international organisations still debate "who should certify AI?" (Evers *et al.*, 2018). Based on the complexity of the landscape and consumers' approval of AI certification entities that emerged from Parts A and B, our final contribution is a research agenda for marketing/social media scholars interested in AI trust/fairness and its certification following the identified concepts in Part C.

## Limitations

In spite of the merits of this work and its quest to introduce AI certification to the marketing and social media communities the novelty of our topic and approach comes with several noteworthy limitations. We will outline these in our respective parts.

Part A used a prospector approach because of the interdisciplinary nature of our literature. We followed Breslin and Gatrell (2023) but encourage a more traditional, structured literature review within the marketing field once the topic has matured. Our strong focus on certifying entities (from a consumer perspective) means that we had to leave important AI certification-relevant areas such as computational methods unincluded.

As our empirical study of Part B mainly served the purpose of providing an example of the use of our developed overview, we allowed for some methodological drawbacks, which we would like to describe further. Firstly, we acknowledge the standard limitations of online crowdworking platforms, such as possible lower attention to the task and no stratified sampling, possibly reducing the quality of the data. However, using a crowd-working platform allows us to capture a more representative sample of the UK compared to a university lab setting (Buhrmester *et al.*, 2011). We opted for Prolific as participants' attention and overall data quality are higher than on other platforms such as MTurk and CrowdFlower (Eyal *et al.*, 2021). Secondly, although we measured participants' approval rates at the start of the questionnaire and opted for a between-subject design to shorten the questionnaire, scenario fatigue may still have occurred because of the amount and novelty of AI decisions. On average, participants spent $M = 8.56$ (SD = 6.36) minutes on the questionnaire. Thirdly, consumers limited knowledge about AI in general (Zhang and Dafoe, 2019) might have hindered their ability to understand differences between the AI entities and the different domains, as indicated by their lower self-reported AI expertise scores

($M = 3.27$, SD = 1.31). Besides addressing the above limitations, we very much believe that a deeper insight into the psychological reasons for consumers' entity preference needs to be established. For that, a qualitative approach and additional empirical studies are interesting future research suggestions.

Part C provides a rich overview of research opportunities for AI certification in marketing and social media. Nonetheless, it can somewhat be seen as "present-oriented" and is currently limited to the state of AI certification. AI technology continually evolves; certification programs designed for today's AI may not fit with the AI of tomorrow. However, some elements of AI certification are relatively durable, such as the concept of fairness, derived from universal ethical principles that apply to any AI technology and society in general (Cihon *et al.*, 2021). Different actors will remain involved in AI over the years (such as consumers, companies and regulating entities), and while we acknowledge the dynamic nature of technology and AI certification, AI certification deserves attention in the marketing field and connects to literature on *consumer trust* and *AI acceptance* (Longoni *et al.*, 2019; Yalcin *et al.*, 2022; Zhu *et al.*, 2022). In addition, trust in AI (certification) will likely play an important role in the larger marketing debate on AI acceptance. Our overview and research agenda provide opportunities to develop these fields further. In spite of these fundamental perspectives on AI certification, we acknowledge that the fast-paced nature of AI (development) might also require an even more futuristic look at the research agenda. To address this, we map out potential research areas for the upcoming years inspired by more mature, regulation-heavy adjunct fields (i.e. financial audit; Hay, 2015). Examples include the certification of more extensive data or even entire AI operations/organisations instead of single AI decisions (Cihon *et al.*, 2021; Shneiderman, 2020; Yanisky-Ravid and Hallisey, 2019). In addition, intensified AI legislation may lead to a forced rotation of certifying entities, which could require the choice of entities within one category and alternative quality control methods (Hay, 2015). While globalisation might lead to an international set of ethical AI standards, we might also observe strong regional quality requirements, with a potential difference between developed and developing economies. There is even the potential for AI programs to detect bias in other AIs, removing human governance altogether (Cihon *et al.*, 2021).

## References

Ada Lovelace Institute, AI Now, and Open Government Partnership (2021), "Algorithmic accountability for the public sector", available at: www.opengovpartnership.org/wp-content/uploads/2021/08/executive-summaryalgorithmic-accountability.pdf (accessed 27 December 2022).

Adamson, A.S. and Smith, A. (2018), "Machine learning and health care disparities in dermatology", *JAMA Dermatology*, Vol. 154 No. 11, pp. 1247-1248.

AI for Good (2023), "Helsinki and Amsterdam launch AI registers to detail city systems", available at: https://aiforgood.itu.int/helsinki-and-amsterdam-launch-ai-registers-to-detailcity-systems/ (accessed 17 August 2022).

Algemene Rekenkamer (2022), "An audit of 9 algorithms used by the dutch government. Report | Netherlands court of audit", available at: https://english.rekenkamer.nl/publications/reports/2022/05/18/an-audit-of-9algorithms-used-by-the-dutch-government (accessed 17 August 2022).

AlgorithmWatch (2021), "Draft AI act: EU needs to live up to its own ambitions in terms of governance and enforcement", available at: https://algorithmwatch.org/en/wpcontent/uploads/2021/08/EU-AI-Act-Consultation-Submission-by-AlgorithmWatch-August-2021.pdf (accessed 15 August 2022).

Arnold, M., Bellamy, R.K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Natesan Ramamurthy, K., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J. and Varshney, K.R. (2019), "FactSheets: increasing trust in AI services through supplier's declarations of conformity", *IBM Journal of Research and Development*, Vol. 63 Nos 4/5, pp. 1-6.

Arora, S.D., Singh, G.P., Chakraborty, A. and Maity, M. (2022), "Polarization and social media: a systematic review and research agenda", *Technological Forecasting and Social Change*, Vol. 183, p. 121942.

Badran, A. (2021), "Thoughts and reflections on the case of Qatar: should artificial intelligence be regulated?", *Artificial Intelligence in the Gulf*, Palgrave Macmillan, Singapore, pp. 69-92.

Baird, A. and Schuller, B. (2020), "Considerations for a more ethical approach to data in AI: on data representation and infrastructure", *Frontiers in Big Data*, Vol. 3, p. 25.

Baker-Brunnbauer, J. (2021), "Management perspective of ethics in artificial intelligence", *AI and Ethics*, Vol. 1 No. 2, pp. 173-181.

Banker, S. and Khetani, S. (2019), "Algorithm overdependence: how the use of algorithmic recommendation systems can increase risks to consumer well-being", *Journal of Public Policy and Marketing*, Vol. 38 No. 4, pp. 500-515.

Belanche, D., Casaló, L.V. and Flavián, C. (2019), "Artificial intelligence in FinTech: understanding Robo-advisors adoption among customer's", *Industrial Management and Data Systems*, Vol. 119 No. 7, pp. 1411-1430.

Bhattacharjee, A., Dana, J. and Baron, J. (2017), "Anti-profit beliefs: how people neglect the societal benefits of profit", *Journal of Personality and Social Psychology*, Vol. 113 No. 5, p. 671.

Bock, D.E., Wolter, J.S. and Ferrell, O.C. (2020), "Artificial intelligence: disrupting what we know about services", *Journal of Services Marketing*, Vol. 34 No. 3, pp. 317-334.

Brach, S., Walsh, G. and Shaw, D. (2018), "Sustainable consumption and third-party certification labels: consumers' perceptions and reactions", *European Management Journal*, Vol. 36 No. 2, pp. 254-265.

Breslin, D. and Gatrell, C. (2023), "Theorizing through literature reviews: the minerprospector continuum", *Organizational Research Methods*, Vol. 26 No. 1, pp. 139-167.

Buhrmester, M., Kwang, T. and Gosling, S.D. (2011), "Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data?", *Perspectives on Psychological Science*, Vol. 6 No. 1, pp. 3-5.

Candelon, F., Evgeniou, T. and Courtaux, M. (2022), "Your business needs an A.I. watchdog: here's how to make sure it has teeth", available at: https://fortune.com/2022/03/04/artificial-intelligence-ai-watchdog-review-board/ (accessed 20 April 2022).

Cao, L. (2018), "AI in finance: a review", *ACM Computing Surveys*, Vol. 50 No. 3, pp. 1-35.

Castelo, N., Bos, M.W. and Lehmann, D.R. (2019), "Task-dependent algorithm aversion", *Journal of Marketing Research*, Vol. 56 No. 5, pp. 809-825.

Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M. and Zhu, H. (2019), "Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders", *Proceedings of the 2019 chi conference on human factors in computing systems*, Glasgow, Scotland, UK, pp. 1-12.

Christensen, T. and Lægreid, P. (2005), "Trust in government: the relative importance of service satisfaction, political factors, and demography", *Public Performance and Management Review*, Vol. 28 No. 4, pp. 487-511.

Cihon, P., Kleinaltenkamp, M.J., Schuett, J. and Baum, S.D. (2021), "AI certification: advancing ethical practice by reducing information asymmetries", *IEEE Transactions on Technology and Society*, Vol. 2 No. 4, pp. 200-209.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. and Starnini, M. (2021), "The echo chamber effect on social media", *Proceedings of the National Academy of Sciences*, Vol. 118 No. 9, p. e2023301118.

Clery, E., Curtice, J., Frankenburg, S., Morgan, H. and Reid, S. (Eds) (2021), *British Social Attitudes: The 38th Report*, Vol. 44, National Centre for Social Research. Consumer motivation, understanding and use, London, Food policy, pp. 177-189.

Cunneen, M., Mullins, M. and Murphy, F. (2019), "Autonomous vehicles and embedded artificial intelligence: the challenges of framing machine driving decisions", *Applied Artificial Intelligence*, Vol. 33 No. 8, pp. 706-731.

Danish Ministery of Industry, Business and Financial Affairs (2019), "New seal for IT security and responsible data use is in its way", available at: https://eng.em.dk/news/2019/oktober/new-seal-for-it-security-and-responsible-data-useis-in-its-way (accessed 13 January 2022).

Dastin, J. (2018), "Amazon scraps secret AI recruiting tool that showed bias against women", available at: www.reuters.com/article/us-amazon-com-jobs-automation-insightidUSKCN1MK08G (accessed 13 January 2022).

Datta, A., Tschantz, M.C. and Datta, A. (2014), "Automated experiments on ad privacy settings: a tale of opacity, choice, and discrimination", arXiv preprint arXiv:1408.6491.

Davenport, T., Guha, A., Grewal, D. and Bressgott, T. (2020), "How artificial intelligence will change the future of marketing", *Journal of the Academy of Marketing Science*, Vol. 48 No. 1, pp. 24-42.

Davies, B., Lalot, F., Peitz, L., Heering, M.S., Ozkececi, H., Babaian, J., Davies, K., Broadwood, J. and Abrams, D. (2021), "Changes in political trust in Britain during the COVID-19 pandemic in 2020: integrated public opinion evidence and implications", *Humanities and Social Sciences Communications*, Vol. 8 No. 1, pp. 1-9.

de Laat, P.B. (2021), "Companies committed to responsible AI: from principles towards implementation and regulation?", *Philosophy and Technology*, Vol. 34 No. 4, pp. 1135-1193.

De Stefano, V. (2019), "'Negotiating the algorithm': automation, artificial intelligence, and labor protection", *Comparative Labor Law & Policy Journal*, Vol. 41, p. 15.

Deloitte (2023), "Trustworthy AITM", available at: www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-aiframework.html (accessed 13 January 2022).

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015), "Algorithm aversion: people erroneously avoid algorithms after seeing them err", *Journal of Experimental Psychology. General*, Vol. 144 No. 1, p. 114, doi: 10.1037/xge0000033.

Erdélyi, O.J. and Goldsmith, J. (2018), "Regulating artificial intelligence: proposal for a global solution", *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, *New Orleans, LA, USA*, pp. 95-101.

European Commission (2021), "Draft AI regulation", available at: https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (accessed 9 September 2021).

European Economic and Social Committee (2019), "The EESC proposes introducing EU certification for "trusted AI" products", available at: www.eesc.europa.eu/en/news-media/news/eesc-proposes-introducing-eucertification-trusted-ai-products (accessed 9 September 2021).

Evers, C., Marchiori, D.R., Junghans, A.F., Cremers, J. and De Ridder, D.T.D. (2018), "Citizen approval of nudging interventions promoting healthy eating: the role of intrusiveness and trustworthiness", *BMC Public Health*, Vol. 18 No. 1, pp. 1-10.

Eyal, P., David, R., Andrew, G., Zak, E. and Ekaterina, D. (2021), "Data quality of platforms and panels for online behavioral research", *Behavior Research Methods*, Vol. 54 No. 4, pp. 1643-1662.

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, E., Goodloe, A., Gupta, J., Hart, C., Jirotka, M., Johnson, H., LaPointe, C., Llorens, A.J., Mackworth, A.K., Maple, C., Pálson, S.M., Pasquale, F., Winfield, A. and Yeong, Z.K. (2021), "Governing AI safety through independent audits", *Nature Machine Intelligence*, Vol. 3 No. 7, pp. 566-571.

Fan, W., Liu, J., Zhu, S. and Pardalos, P.M. (2020), "Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS)", *Annals of Operations Research*, Vol. 294 Nos 1/2, pp. 567-592.

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

Garcia-Garcia, M., Wahren, B., Midkiff, M. and Espinosa, E. (2021), "Forget statements: consumers want deeper social-justice commitment from brand", available at: www.ipsos.com/sites/default/files/ct/publication/documents/2021-04/21-04-53_Forget_pov_v3.pdf (accessed 23 December 2022).

Gillespie, N., Lockey, S. and Curtis, C. (2021), *Trust in Artificial Intelligence: A Five Country Study*, The University of Queensland and KPMG Australia, doi: 10.14264/e34bfa3.

Google, A.I. (2023), "Our principles", Link.

Grunert, K.G., Hieke, S. and Wills, J. (2014), "Sustainability labels on food products: consumer motivation, understanding and use", *Food Policy*, Vol. 44, pp. 177-189.

Guihot, M., Matthew, A.F. and Suzor, N.P. (2017), "Nudging robots: innovative solutions to regulate artificial intelligence", *Vanderbilt Journal of Entertainment and Technology Law*, Vol. 20, p. 385.

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M. and Katyal, V. (2018), "Why We need to audit algorithms", *Harvard Business Review*, available at: https://hbr.org/2018/11/whywe-need-to-audit-algorithms (accessed 13 January 2022).

Haleem, A., Javaid, M., Qadri, M.A., Singh, R.P. and Suman, R. (2022), "Artificial intelligence (AI) applications for marketing: a literature-based study", *International Journal of Intelligent Networks*, Vol. 3, pp. 119-132.

Hasan, R., Shams, R. and Rahman, M. (2021), "Consumer trust and perceived risk for voicecontrolled artificial intelligence: the case of Siri", *Journal of Business Research*, Vol. 131, pp. 591-597.

Hay, D. (2015), "The frontiers of auditing research", *Meditari Accountancy Research*, Vol. 23 No. 2, pp. 158-174.

Hermann, E. (2021), "Artificial intelligence and mass personalization of communication content—an ethical and literacy perspective: new media & society", *New Media and Society*, Vol. 24 No. 5, pp. 1258-1277, doi: 10.1177/14614448211022702.

Howard, A. and Borenstein, J. (2018), "The ugly truth about ourselves and our robot creations: the problem of bias and social inequity", *Science and Engineering Ethics*, Vol. 24 No. 5, p. 15211536.

Hufnagel, G., Schwaiger, M. and Weritz, L. (2022), "Seeking the perfect price: consumer responses to personalized price discrimination in e-commerce", *Journal of Business Research*, Vol. 143, pp. 346-365.

I Amsterdam (2020), "Amsterdam launches AI algorithm registry", available at: www.iamsterdam.com/en/business/news-and-insights/news/2020/amsterdamlaunches-ai-algorithm-registry (accessed 5 January 2022).

IBM (2019), "Everyday ethics for artificial intelligence", available at: www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf (accessed 23 December 2022).

IPSOS (2022), "Global opinions and expectations about AI 2022", available at: www.ipsos.com/sites/default/files/ct/news/documents/2022-01/Globalopinions-and-expectations-about-AI-2022.pdf (accessed 23 December 2022).

Janssen, M. and Hamm, U. (2012), "Product labelling in the market for organic food: consumer preferences and willingness-to-pay for different organic certification logos", *Food Quality and Preference*, Vol. 25 No. 1, pp. 9-22.

Jobin, A., Ienca, M. and Vayena, E. (2019), "The global landscape of AI ethics guidelines", *Nature Machine Intelligence*, Vol. 1 No. 9, pp. 389-399.

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

Kay, M., Matuszek, C. and Munson, S.A. (2015), "Unequal representation and gender stereotypes in image search results for occupations", *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 3819-3828.

Kaya, F., Aydin, F., Schepman, A., Rodway, P., Yetişensoy, O. and Demir Kaya, M. (2022), "The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence", *International Journal of Human–Computer Interaction*, pp. 1-18.

Kim, K. and Kim, J. (2011), "Third-party privacy certification as an online advertising strategy: an investigation of the factors affecting the relationship between third-party certification and initial trust", *Journal of Interactive Marketing*, Vol. 25 No. 3, pp. 145-158.

Kim, J., Giroux, M. and Lee, J.C. (2021), "When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations", *Psychology and Marketing*, Vol. 38 No. 7, pp. 1140-1155.

Knowles, B. and Richards, J.T. (2021), "The sanction of authority: promoting public trust in AI", Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 262-271.

Konuk, F.A. (2019), "Consumers' willingness to buy and willingness to pay for fair trade food: the influence of consciousness for fair consumption, environmental concern, trust and innovativeness", *Food Research International*, Vol. 120, pp. 141-147.

KPMG (2023), "Controlling AI", available at: https://advisory.kpmg.us/articles/2019/controlling-ai.html (accessed 20 April 2022).

Lambrecht, A. and Tucker, C. (2019), "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads", *Management Science*, Vol. 65 No. 7, pp. 2966-2981.

Lanero, A., Vázquez, J.L. and Sahelices-Pinto, C. (2021), "Halo effect and source credibility in the evaluation of food products identified by third-party certified Eco-Labels: can information prevent biased inferences?", *Foods*, Vol. 10 No. 11, p. 2512.

Lilkov, D. (2021), "Regulating artificial intelligence in the EU: a risky game", *European View*, Vol. 20 No. 2, pp. 166-174.

Lima, G. and Cha, M. (2021), "Descriptive AI ethics: collecting and understanding the public opinion", arXiv preprint arXiv:2101.05957.

Long, D. and Magerko, B. (2020), "What is AI literacy? Competencies and design considerations", *Proceedings of the 2020 CHI conference on human factors in computing system, Honolulu, HI, USA*, pp. 1-16.

Longoni, C., Bonezzi, A. and Morewedge, C.K. (2019), "Resistance to medical artificial intelligence", *Journal of Consumer Research*, Vol. 46 No. 4, pp. 629-650.

McKinsey (2021), "The state of AI in 2021", available at: www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Global%20survey%20The%20state%20of%20AI%20in%202021/Global-survey-The-state-of-AI-in-2021.pdf (accessed 2 May 2023).

Mac, R. (2021), "Facebook apologizes after A.I. Puts' primates' label on video of black men", available at: https://advisory.kpmg.us/articles/2019/controlling-ai.html (accessed 27 April 2022).

Malodia, S., Ferraris, A., Sakashita, M., Dhir, A. and Gavurova, B. (2022), "Can Alexa serve customers better? AI-driven voice assistant service interactions", *Journal of Services Marketing*, Vol. 37 No. 1, pp. 25-39.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021), "A survey on bias and fairness in machine learning", *ACM Computing Surveys*, Vol. 54 No. 6, pp. 1-35.

Metaxa, D., Park, J.S., Robertson, R.E., Karahalios, K., Wilson, C., Hancock, J. and Sandvig, C. (2021), "Auditing algorithms: understanding algorithmic systems from the outside in. Foundations and trends® in human–computer interaction", *Foundations and Trends® in Human–Computer Interaction*, Vol. 14 No. 4, pp. 272-344.

Microsoft (2018), "The future computed – artificial intelligence and its role in society", available at: https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificialintelligence-role-society/ (accessed 23 December 2022).

Mökander, J., Axente, M., Casolari, F. and Floridi, L. (2022), "Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation", *Minds and Machines*, Vol. 32 No. 2, pp. 241-268.

Moss, E., Watkins, E.A., Singh, R., Elish, M.C. and Metcalf, J. (2021), "Assembling accountability", available at: https://datasociety.net/wpcontent/uploads/2021/06/Assembling-Accountability.pdf (accessed 8 August 2022).

Nawaz, T. (2018), "Determinants and consequences of disruptive innovations: evidence from the UK financial services sector", *Journal of Accounting and Management Information Systems*, Vol. 17 No. 2, pp. 234-251.

O'Brien, C. (2020), "Facebook civil rights audit urges 'mandatory' algorithmic bias detection", available at: https://venturebeat.com/ai/facebook-civil-rights-audit-urgesmandatory-algorithmic-bias-detection/ (accessed 25 December 2022).

OECD Data (2022), "Trust in government, OECD", available at: https://data.oecd.org/gga/trustin-government.htm (accessed 9 January 2021).

OECD Legal Instruments (2019), "Recommendation of the council on artificial intelligence", available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed 19 January 2022).

O'Neil, C. (2017), *Weapons of Math Destruction*, Penguin Books, Harlow.

O'Shaughnessy, M., Schiff, D., Varshney, L.R., Rozell, C. and Davenport, M. (2021), *What Governs Attitudes toward Artificial Intelligence Adoption and Governance*?, GA Institute of Technology, University of IL at Urbana-Champaign, Atlanta, IL, 14 December.

Oxford Insights (2019), "Government artificial intelligence readiness index", available at: https://africa.ai4d.ai/wp-content/uploads/2019/05/ai-gov-readiness-report_v08.pdf (accessed 1 September 2021).

Pandey, A. and Caliskan, A. (2021), "Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms", *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 822-833.

Park, C.W., Mothersbaugh, D.L. and Feick, L. (1994), "Consumer knowledge assessment", *Journal of Consumer Research*, Vol. 21 No. 1, pp. 71-82.

Park, S.K., Yang, Y. and Zhang, S. (2022), "Mitigating inequalities caused by awareness of algorithmic bias", Working paper.

Philip, R.E. (2021), "AI and IoT can help to make smart home", available at: https://medium.com/@rohithaelsa/how-ai-and-iot-can-help-to-make-smart-home-5133d050505f (accessed 20 December 2022).

Pirson, M., Martin, K. and Parmar, B. (2019), "Public trust in business and its determinants", *Business and Society*, Vol. 58 No. 1, pp. 132-166.

Prahl, A. and Goh, W.W.P. (2021), "'Rogue machines' and crisis communication: when AI fails, how do companies publicly respond?", *Public Relations Review*, Vol. 47 No. 4, p. 102077, doi: 10.1016/j.pubrev.2021.102077.

Puntoni, S., Reczek, R.W., Giesler, M. and Botti, S. (2021), "Consumers and artificial intelligence: an experiential perspective", *Journal of Marketing*, Vol. 85 No. 1, pp. 131-151.

PwC (2018), "AI predictions", available at: www.pwc.es/es/home/assets/aipredictions-2018-report.pdf (accessed 10 October 2020).

Rahwan, I. (2018), "Society-in-the-loop: programming the algorithmic social contract", *Ethics and Information Technology*, Vol. 20 No. 1, pp. 5-14.

Raji, I.D., Xu, P., Honigsberg, C. and Ho, D. (2022), "Outsider oversight: designing a third party audit ecosystem for AI governance", *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 557-571.

Roe, B. and Teisl, M.F. (2007), "Genetically modified food labeling: the impacts of message and messenger on consumer perceptions of labels and products", *Food Policy*, Vol. 32 No. 1, pp. 49-66.

Roski, J., Maier, E.J., Vigilante, K., Kane, E.A. and Matheny, M.E. (2021), "Enhancing trust in AI through industry self-governance", *Journal of the American Medical Informatics Association*, Vol. 28 No. 7, pp. 1582-1590.

Scherer, M.U. (2015), "Regulating artificial intelligence systems: risks, challenges, competencies, and strategies", *Harvard Journal of Law and Technology*, Vol. 29, p. 353.

Sharkov, G., Todorova, C. and Varbanov, P. (2021), "Strategies, policies, and standards in the EU towards a roadmap for robust and trustworthy AI certification", *Information and Security: An International Journal*, Vol. 50 No. 1, pp. 11-22.

Shin, D. (2020), "How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance", *Computers in Human Behavior*, Vol. 109, p. 106344.

Shneiderman, B. (2020), "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI", *Systems. ACM Transactions on Interactive Intelligent Systems (TiiS)*, Vol. 10 No. 4, pp. 1-31.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxeninan, A.A., Shah, J., Tambe, M. and Teller, A. (2016), "Artificial 'intelligence and life in 2030'", available at: https://ai10020201023.sites.stanford.edu/sites/g/files/sbiybj18871/files/media/file/ai100report10032016fnl_singles.pdf (accessed 10 October 2020).

Stuurman, K. and Lachaud, E. (2022), "Regulating AI. A label to complete the proposed act on artificial intelligenc", *e. Computer Law and Security Review*, Vol. 44, p. 105657.

Sushina, T. and Sobenin, A. (2020), "Artificial intelligence in the criminal justice system: leading trends and possibilities", *Proceedings of the, 6th International Conference on Social, economic, and academic leadership (ICSEAL-6-2019)*. Atlantis Press, Prague, Czech Republic.

Taeihagh, A. (2021), "Governance of artificial intelligence", *Policy and Society*, Vol. 40 No. 2, pp. 137-157.

Tutt, A. (2017), "An FDA for algorithms", *Administrative Law Review*, Vol. 69, p. 83.

UK Government (2023), "Departments, agencies and public bodies", available at: www.gov.uk/government/organisations (accessed 23 December 2022).

Valand, D. (2021), "How Netflix uses AI for content creation and recommendation", available at: https://medium.com/swlh/how-netflix-uses-ai-for-content-creation-andrecommendation-c1919efc0af4 (accessed 23 December 2022).

Veale, M. and Borgesius, F.Z. (2021), "Demystifying the draft EU artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach", *Computer Law Review International*, Vol. 22 No. 4, pp. 97-112.

Vigdor, N. (2019), "Apple card investigated after gender discrimination complaints", available at: www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html (accessed 23 December 2022).

Vincent, J. (2019), "Google and Microsoft warn investors that bad AI could harm their brand", The Verge, Link accessed 29 July, 2022.

Vlačić, B., Corbo, L.E., Silva, S.C. and Dabić, M. (2021), "The evolving role of artificial intelligence in marketing: a review and research agenda", *Journal of Business Research*, Vol. 128, pp. 187-203.

Wang, R., Harper, F.M. and Zhu, H. (2020), "Factors influencing perceived fairness in algorithmic decision-making: algorithm outcomes, development procedures, and individual

differences", Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-14.

Winfield, A.F. and Jirotka, M. (2018), "Ethical governance is essential to building trust in robotics and artificial intelligence systems", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 376 No. 2133, p. 20180085.

Winter, P.M., Eder, S., Weissenböck, J., Schwald, C., Doms, T., Vogt, T., … and Nessler, B. (2021), "Trusted artificial intelligence: towards certification of machine learning applications", arXiv preprint arXiv:2103.16910.

Wright, J., Leslie, D., Raab, C., Kitagawa, F., Ostmann, F. and Briggs, M. (2021), "Privacy, agency and trust in human-AI ecosystems: interim report (short version)", in *The Alan Turing Institute*.

Yalcin, G., Lim, S., Puntoni, S. and van Osselaer, S.M. (2022), "Thumbs up or down: consumer reactions to decisions by algorithms versus humans", *Journal of Marketing Research*, Vol. 59 No. 4, p. 222437211070016.

Yang, Q. and Tang, W. (2010), "Exploring the sources of institutional trust in China: culture, mobilization, or performance?", *Asian Politics and Policy*, Vol. 2 No. 3, pp. 415-436.

Yanisky-Ravid, S. and Hallisey, S.K. (2019), "Equality and privacy by design: a new model of artificial intelligence data transparency via auditing, certification, and SafeHarbor regimes", *Fordham Urban Law Journal*, Vol. 46, p. 42.

Zhang, B. and Dafoe, A. (2019), "Artificial intelligence: American attitudes and trends", available at: https://isps.yale.edu/sites/default/files/files/Zhang_us_public_opinion_report_jan_2019.pdf (accessed 10 October 2020).

Zhang, S., Mehta, N., Singh, P.V. and Srinivasan, K. (2021), "Can an AI algorithm mitigate racial economic inequality? An analysis in the context of Airbnb: an analysis in the context of Airbnb", Working paper (3770371), Rotman School of Management, 21 Januari.

Zhu, Y., Zhang, J., Wu, J. and Liu, Y. (2022), "AI is better when I'm sure: the influence of certainty of needs on consumers' acceptance of AI chatbots", *Journal of Business Research*, Vol. 150, pp. 642-652.

Zuloaga, L. (2021), "Industry leadership: new audit results and decision on visual analysis", available at: www.hirevue.com/blog/hiring/industry-leadership-new-auditresults-and-decision-on-visual-analysis (accessed 8 August 2022).

## Further reading

AI Incidents Database (2022), available at: https://incidentdatabase.ai (accessed 23 December 2022).

Fair, A.I. (2023), "Our vision", available at: www.fairai.uk/our-vision (accessed 13 January 2022).

## Appendix

**Table A1.** Approval means and standard deviations per entity per AI decision

| | AI decision | Government | NGO | Commercial third-party | Self-declaration |
|---|---|---|---|---|---|
| Social | Social housing allocation | $M = 4.33$ (SD = 1.87) | $M = 4.16$ (SD = 1.88) | $M = 4.01$ (SD = 2.01) | $M = 4.06$ (SD = 1.97) |
| | Pension allocation | $M = 4.58$ (SD = 1.75) | $M = 4.04$ (SD = 1.79) | $M = 4.00$ (SD = 1.97) | $M = 3.92$ (SD = 1.97) |
| Safety | Predictive crime policing | $M = 5.18$ (SD = 1.55) | $M = 4.58$ (SD = 1.82) | $M = 4.15$ (SD = 2.02) | $M = 4.11$ (SD = 2.10) |
| | Real-time analysis of safety cameras | $M = 5.20$ (SD = 1.52) | $M = 4.81$ (SD = 1.77) | $M = 4.26$ (SD = 1.84 | $M = 4.33$ (SD = 2.04) |
| Health | Imaging diagnostics | $M = 4.90$ (SD = 1.68) | $M = 4.80$ (SD = 1.73) | $M = 4.50$ (SD = 1.80) | $M = 4.29$ (SD = 1.92) |
| | Public health prediction | $M = 5.25$ (SD = 1.52) | $M = 4.79$ (SD = 1.69) | $M = 4.23$ (SD = 2.01) | $M = 4.43$ (SD = 1.93) |
| Finance | Price of insurance | $M = 4.51$ (SD = 1.73) | $M = 4.40$ (SD = 1.70) | $M = 4.14$ (SD = 1.96) | $M = 4.21$ (SD = 1.94) |
| | Height of mortgage | $M = 4.40$ (SD = 1.80) | $M = 4.21$ (SD = 1.79) | $M = 4.25$ (SD = 1.86) | $M = 4.00$ (SD = 2.00) |

### Supplementary material

The supplementary material for this article can be found online.

### Corresponding author

Myrthe Blösser can be contacted at: m.blosser@uva.nl