Awareness of Unethical Artificial Intelligence and its Mitigation Measures

Sonja HÖLLER
MCI Management Center Innsbruck, Austria
s.hoeller@mci4me.at

Thomas DILGER
MCI Management Center Innsbruck, Austria
thomas.dilger@mci.edu

Teresa SPIESS
MCI Management Center Innsbruck, Austria
teresa.spiess@mci.edu

Christian PLODER
MCI Management Center Innsbruck, Austria
christian.ploder@mci.edu

Reinhard BERNSTEINER
MCI Management Center Innsbruck, Austria
reinhard.bernsteiner@mci.edu

Abstract
The infrastructure of the Internet is based on algorithms that enable the use of search engines, social networks, and much more. Algorithms themselves may vary in functionality, but many of them have the potential to reinforce, accentuate, and systematize age-old prejudices, biases, and implicit assumptions of society. Awareness of algorithms thus becomes an issue of agency, public life, and democracy. Nonetheless, as research showed, people are lacking algorithm awareness. Therefore, this paper aims to investigate the extent to which people are aware of unethical artificial intelligence and what actions they can take against it (mitigation measures). A survey addressing these factors yielded 291 valid responses. To examine the data and the relationship between the constructs in the model, partial least square structural modeling (PLS-SEM) was applied using the Smart PLS 3 tool. The empirical results demonstrate that awareness of mitigation measures is influenced by the self-efficacy of the user. However, trust in the algorithmic platform has no significant influence. In addition, the explainability of an algorithmic platform has a significant influence on the user's self-efficacy and should therefore be considered when setting up the platform. The most frequently mentioned mitigation measures by survey participants are laws and regulations, various types of algorithm audits, and education and training. This work thus provides new empirical insights for researchers and practitioners in the field of ethical artificial intelligence.

Keywords: artificial intelligence; biased artificial intelligence; algorithmic fairness; IT-audit; ethical AI;

JEL Classification: M00; D83; C30

DOI: https://doi.org/10.24818/ejis.2023.17

1. Introduction

Algorithms have become critical to the infrastructure of the Internet, whether through search engines, social networks, or music streaming services. They are technological recipes and logical instructions developed by data engineers, mathematicians, and programmers. To

Received: 1 August 2023; Revised: 9 October 2023; Accepted: 1 November 2023

Internet users, however, they are invisible actors that give them instructions that they consciously or unconsciously follow. As algorithms become more influential in the distribution of information and content, users' awareness of algorithms becomes a matter of "agency, public life, and democracy" (Gran et al., 2021, p. 1).

As research has shown, many people are unaware of the use of algorithms in their daily activities (Gran et al., 2021; Powers, 2017; Rader & Gray, 2015; Shin et al., 2022). However, as societies become increasingly dependent on algorithms, our enduring biases, prejudices, and underlying assumptions are reflected back in digital form through the algorithmic systems we use. As such, they have the capacity to significantly amplify, magnify, and systematize biases while appearing to be objective, neutral arbiters (Rovatsos et al., 2019). This trend is exacerbated by the extraordinary pace of adoption of artificial intelligence (AI) systems by corporations, nonprofits, and governments, which can scale production massively through increased access to artificial intelligence development tools and internet-sourced datasets. There are legitimate concerns about the effectiveness of these automated systems for the full range of users. In particular, the ability of the system to reproduce, reinforce, or exaggerate undesirable current societal biases (Raji et al., 2020). To actively reduce the existence of unethical AI applications, several approaches can be implemented. An important tool against unethical AI is awareness and critical reflection, strengthened by transparency and knowledge, which is a powerful tool for society. With user awareness, behavior can be shaped (Gran et al., 2021). It can also influence demand, leading to the design of user-controllable artificial intelligence (Shin et al., 2022).

Other mitigation measures for unethical AI can be algorithm audit and algorithmic impact assessment. These two approaches are introduced to evaluate algorithms either before or after their deployment. This type of evaluation can be used to anticipate likely harmful consequences before they occur. In addition, they provide decision support for building mitigation strategies, explicitly defining and tracking potential adverse effects, and anticipating negative feedback loops and system-level hazards (Ada Lovelace Institute, 2020; Raji et al., 2020).

Therefore, the purpose of this paper is to measure awareness of unethical artificial intelligence, as well as awareness and knowledge of potential mitigations against discriminatory behavior.

2. Awareness of AI

Algorithms offer users many benefits, but their integration, especially for automated customization, raises moral and privacy issues (Ashok et al., 2022). Algorithms track user activity and regulate content, all behind the scenes (Dwivedi et al., 2021). Common examples are recommendation algorithms, such as those used by Netflix, or filtering algorithms in social media feeds. It is worth noting that algorithms not only enable the flow of content, they actively influence it and in some cases distort it with built-in biases (Gran et al., 2021).

The majority of users do not know how and to what extent the algorithm affects their lives, let alone fully understand what the algorithm does. Several studies have confirmed this lack of awareness. For example, Hamilton et al. (2014) research has shown that only very sophisticated social media users are aware of the influence of algorithms on their content presentation interfaces. However, Rader & Gray asked Amazon Mechanical Turkers, who tend to be more Internet savvy than the average user, whether or not they believe they see every Facebook post their friends make. The result showed that 73% did not believe it (Rader & Gray, 2015). An

interview study asking 40 Facebook users about their knowledge of the Facebook News Feed curation algorithm found that more than half were unaware of its existence (Eslami et al., 2015). While researching the algorithm awareness of the net generation in the US Powers (2017) concluded that they were unaware of the impact of the personalization technology. A recent and nationwide study was conducted by Gran et al. (2021) in Norway, a highly digitized country where almost everyone has access to the Internet and a smartphone. The study concluded that there are different levels of awareness of algorithms according to gender, geography, age and education. However, they concluded that 61% of the Norwegian population reported no or low awareness of algorithms. The groups with the highest awareness were well educated, predominantly male and living in urban areas (Gran et al., 2021). Shin et al. (2022) conducted their research with students at a public university in South Korea who were enrolled in classes on topics such as digital media, algorithms, and UX design. The results of their quantitative research supported previous assumptions about personalized algorithms. The first is based on the observations of Hargittai et al. (2020), who found that users lack basic knowledge about how algorithms work. The second supports Eslami et al.'s examination of the black-box nature of algorithms in everyday AI interactions (Eslami et al., 2015). The last confirmed assumption is that users' insufficient understanding of algorithms prevents them from making wise privacy choices (Shin, 2021; Swart, 2021). In addition, the study itself linked privacy issues to users' awareness and disclosure of information. The results show that self-efficacy is related to self-disclosure, which is influenced by privacy and trust. This shows that it is very important to evaluate algorithm awareness in different nations, societies and educational levels (Shin et al., 2022).

2.1. Consequences from the lack of Algorithm Awareness

The lack of such awareness can have societal consequences that affect public participation and democracy. There is an increased risk of reinforcing democratic deficits by using an algorithmic framework that uses machine learning to automatically reinforce existing patterns. This weakens the foundation for an informed public and democratic engagement. The digital divide of algorithmic awareness will widen with this infrastructure in place. Furthermore, access to information will be actively shaped by the user and amplified by the implemented "smart" structure. Users will become "prosumers" of the infrastructure, i.e. both producers and consumers. An important question to be addressed is "How much can the user be held responsible for - in a you-get-the-infrastructure-you-deserve logic?" (Gran et al., 2021, p. 15).

Gran et al. (2021) argue that education will become crucial for online skills, use and benefits. Other factors influencing algorithmic awareness, like age or gender, will only be temporary. However, this could mean that the division, may it be digital or societal, will increase, plus societal inequalities will be reproduced and accelerated in unforeseen and unanticipated ways. Park & Humphry (2019) predict that new divisions will emerge as a result of the unequal distribution of information and expertise between those who have the ability to question the process of datafication and those who do not. Based on these predictions, it is very important to further explore how social justice is affected and changed by data-driven arrangements (Gran et al., 2021).

2.2. Algorithm Awareness, FEAT Factors, Self-efficacy and IT experts

Addressing the lack of awareness must become a fundamental aspect of algorithmic procedures and use (Eslami et al., 2016). The realization that people do deserve to understand how algorithms work has sparked interest in algorithm awareness (AA) (Sandvig et al., 2014).

To comply with the General Data Protection Regulation (GDPR), user awareness is a prerequisite for algorithmic codes of practice (Shin, 2021). User awareness of algorithms is the process of informing about the existing algorithmic system. Because being aware of the practices can influence the information sharing behavior of users (Zarouali et al., 2021).

As a result, factors related to AA have received more public and research attention. These factors are Fairness, Explanability, Accountability and Transparency, also known as FEAT. AA means that the user is informed about the algorithm used and how it works. It is intertwined with the goal that users can consciously decide whether to use the algorithm platform and what personal information to share with the company behind it. If this is not the case, users can be negatively affected in a variety of ways without realizing it (Shin et al., 2022). Examples vary widely, but they all cause real harm. They include systems that have wrongly denied welfare benefits, kidney transplants, mortgages, or wrongful arrests due to biased facial recognition technology (Costanza-Chock et al., 2022).

Fairness in the context of algorithmic decisions means not constructing discriminatory or unfair consequences. It also means creating reasonable and equitable treatments that conform to accepted rules or principles (Diakopoulos, 2015).

When describing the user's decision making, the term explainability is used (Shin et al., 2022). Helping users understand how the AI works and behaves increases their trust (Rai, 2020). When the system’s workings are understood by the users, they are more likely to use the system in the proper way and develop a relationship of trust (Renjith et al., 2020; Shin & Park, 2019).

The principle of accountability is mainly related to the liability of the creators and providers of the algorithm-driven platform. If the AI platform causes any damage, the responsible people will be liable for the consequences (Diakopoulos, 2015).

The last term, transparency, challenges the description of service reasoning and different types of data usage where sensitive data is involved (Shin & Park, 2019). It requires that the judgments made by algorithms be transparent, unambiguous, verifiable, and/or observable to the users who use, adopt, and are affected by the systems (Shin et al., 2022). This term refers to the black box nature of algorithms, a phenomenon where people do not understand the inner workings of the algorithm because the information is proprietary or too complex. Even though algorithms are complex in their technological operation, people are eager to understand the inner workings of the algorithm, also known as the black box. They want to know what data is being collected and how the input data affects the outputs. There is an expectation of visibility and opacity in the algorithmic processes. This motivation is also supported by the right to explanation of the EU General Data Protection Regulation (Shin & Park, 2019).

Social justice is also influenced by the actions that citizens can take. This concept is described by the term "self-efficacy". It is understood as the user's impression of their own ability to perform a certain level of performance, as well as their ability to handle and accomplish a specific task (Hu et al., 2021). This term is important in the context of algorithmic awareness, as it indicates that the user feels empowered and knowledgeable enough to navigate the platform's user interface. High self-efficacy indicates that the user is interested in the domain. Studies show that it is crucial for motivation and learning, especially because challenges and failures can be overcome (Fryer et al., 2020). In other AI-related studies, measures of self-efficacy have included the ability to carry out plans or tasks on the platforms, to complete planned actions in difficult situations or under time pressure, and to work effectively on multiple tasks at once (Shin et al., 2022). Self-efficacy is known to influence the trust of the

users (Fazal-E-Hasan et al., 2020). It has been researched that a lot of information significantly improves the built trust (Hu et al., 2021).

Experts in a field or experienced decision makers are less likely to have their decisions influenced by bias because their experience and expertise are assumed to improve their decision making. However, there is a shift in the literature on this phenomenon. Some believe that expertise can eliminate or mitigate biased decisions (Bazerman & Moore, 2012), although others argue that the biases, which are fundamental and judgmental in nature, are unlikely to be eliminated over time by experience alone. Correcting them would require precise and immediate feedback, which is unlikely to be available in real-world scenarios (Bereby-Meyer & Grosskopf, 2008). In addition to identifying and eliminating bias, it is also important to make AI explainable and understandable. Most often, AI systems are used to help less experienced users make complex decisions. Paradoxically, without the necessary domain expertise, it is very difficult to interpret the results. Research has shown that less experienced users trust or rely on the AI system more than the experts. For example, Micocci et al. (2021) investigated that less experienced physicians were more likely to accept incorrect AI recommendations. Therefore, the awareness that the AI system may make mistakes needs to be conveyed to the users. In general, human users should provide a certain amount of knowledge to be able to compensate for errors made by the AI (Zhang et al., 2020). Since this factor influences the awareness of unethical AI and its mitigation measures, the questionnaire included a question about years of experience in the IT field. This will test whether there is a difference in awareness and knowledge of mitigation measures between the different experience groups.

3. Mitigation Measures

With the current rise of artificial intelligence, numerous new applications have been developed in the economic, scientific, and artistic fields. However, with this widespread application came a growing awareness of the impact of AI systems and a realization that current industry and academic norms are insufficient to ensure responsible AI development. The AI community has responded to this realization and attempted to address it, resulting in the adoption of ethical principles by research and technology companies. However, these principles are not legally enforceable, and translating them into action is often not obvious. In addition, it is usually not possible to evaluate the actions of AI developers in terms of their ethical practices because the code is not public. Furthermore, it is not possible to hold developers accountable for any deviations from the principles or behaviors, leading to accusations of "ethics washing" (Brundage et al., 2020).

This leads to mistrust of the AI system, which should be rebuilt to build responsible AI. This could be done through mechanisms to demonstrate responsible behavior and to verify ethical claims (Brundage et al., 2020). Therefore, different types of mechanisms are defined, which will be explained in more detail below.

3.1. Algorithmic Audit

Although each algorithm is used in a different environment, with different objectives and uses, each algorithm can be audited. However, the applicability of an algorithmic audit is linked to certain requirements, such as the algorithm itself, the context, the responsible persons, and the administrative and legal environment.

For algorithmic audits, the main focus is on algorithms that have a social impact. In this context, social impact means the collection and processing of personal or sensitive data and the intrusion into the lives of individuals as well as into the lives of important social groups or vulnerable groups. These impacts can be either positive or negative. If it is considered negative, some kind of bias or discrimination is implemented. In this way, the algorithm can reproduce or reinforce existing inequalities or create new ones that harm vulnerable people and groups. They can also be linked to a violation of personal data protection and privacy rights. Algorithmic audits make it possible to make the technology used more understandable, transparent, predictable, and subject to control by the public, government agencies, and corporations. This can be done before the system is developed, during its development, or a posteriori (Eticas Consulting, 2021; Raji et al., 2020). During its research, the Ada Lovelace Institute (Ada Lovelace Institute, 2020) discovered that two different meanings of audit are commonly used. The first is from the perspective of the computer science community, where the social science practice of auditing is applied to algorithmic systems. This means that a particular hypothesis is tested by a narrowly focused test on a system, looking at its inputs and outputs. For example, one might want to test whether racial bias is embedded in the outcomes of decisions. The Ada Lovelace Institute calls this type of audit a bias audit. They are typically conducted by external and independent actors who do not have the benefit of working closely with the team or organization that designed and deployed the AI system. These audits are typically done on systems in use (Ada Lovelace Institute, 2020). So far, independent researchers or journalists have mostly used bias audits, although this would be a practical way for developers to test their own systems. The techniques used for bias audits can vary, but the core is that only one hypothesis is tested. Therefore, each test can only examine one metric, such as race, gender, or age, which is why these methods cannot provide a holistic view of the system. Furthermore, just because a bias audit would not conclude that the system is wretched, it could be that other forms of discrimination, ethical issues, or negative impacts on society are occurring elsewhere in the system (Ada Lovelace Institute, 2020).

The second type comes from the common understanding of audit, where it means a comprehensive inspection and compliance exercise, similar to a financial audit. In this sense, an audit performs a comprehensive inspection to analyze whether an algorithmic system acts according to its rules or standards. This type is called a regulatory inspection: it tests the entire lifecycle of a system against specific regulations, such as data protection laws, equality legislation, or insurance industry requirements. This detailed analysis requires the cooperation of the people who use the algorithmic system. This type of inspection is usually carried out by regulators with statutory powers or by auditors. They may analyze an entire product, model, or algorithm, examining the code, inputs, outputs, and documentation, as well as the organizational processes and human behavior around it. The exact scope of a regulatory inspection depends on the context, making it difficult to establish a standardized approach. Different tools can be used during the inspection, such as bias auditing, mandating data access, inspecting the operation of the system, talking to developers and users, and inspecting the code of the AI system. Interest in these types of inspections is growing, but real-world examples are still few and far between. This is due to the newness of the field. Auditing firms need to retrain and establish standards for this type of inspection (Ada Lovelace Institute, 2020).

The Supreme Audit Institutions (SAIs) of Brazil, Finland, Germany, the Netherlands, Norway, and the United Kingdom joined forces to establish methodologies and practices to enable successful and efficient audits. The SAIs of Germany, the Netherlands, Norway, and the United Kingdom established a catalogue for auditing the compliance and performance of machine learning algorithms. Their audit areas include understanding the data, the model development

process, the performance of the model, and ethical considerations such as accountability and fairness (Supreme Audit Institutions of Finland Germany the Netherlands Norway and UK, 2020). The SAI paper emphasizes the need for independent third-party auditing to ensure the security of fundamental rights. Research on auditing machine learning algorithms is ongoing, which means that the SAIs' paper will be adapted as new findings emerge from their pilot projects (Supreme Audit Institutions of Finland Germany the Netherlands Norway and UK, 2020).

In general, algorithmic audits have made a major contribution to improving the attribution of responsibility and accountability of algorithmic systems. Their introduction contributes significantly to raising public awareness of algorithmic accountability (Eticas Consulting, 2021; Raji et al., 2020).

3.2. Algorithmic Impact Assessments

Algorithmic impact assessment is the second group of studies of AI systems. The Ada Lovelace Institute (Ada Lovelace Institute, 2020) categorizes them into two fields. The first one is the *algorithmic risk assessment*. It is used before the system is deployed and analyzes the possibilities of impact area and the risks that come with it. It is commonly used for analyzing the environmental impacts, data protection impacts or risks focused on particular harms. Additionally, Mantelero (2018) created an assessment including human-rights, ethics and the social impact (HRESIA). It is designed to combine the protection of personal data and the fundamental rights and freedoms of individuals, regardless of their geographical location (Mantelero, 2018). At present, they are mainly used in the public sector. Their main application is before the system is deployed, but algorithmic risk assessment can also be used as a continuous method to assess changing risks or characteristics. There is no clear consensus on their requirements and application, or on what constitutes best practice (Ada Lovelace Institute, 2020).

The second is *algorithmic impact evaluations*. They are conducted after the system has been installed. Their purpose is to assess the impact of the system on a given population, for example, to identify political or economic influences. Therefore, they are mainly used by researchers in the public sector, but also in the private sector, analyzing data that was not available before the system was implemented. Other applications in the technology sector include human rights impact assessments, which assess the impact of business projects on human rights defenders. However, it is unclear whether the developer is responsible for implementing the recommendations of the assessments (Ada Lovelace Institute, 2020).

3.3. Other Mitigation Measures

As mentioned in section 2, the goal should be to establish trustworthy AI, which can be achieved by fulfilling the requirements mentioned. This can be done by implementing technical and non-technical methods (European Commision’s High-Level Expert Group on Artificial Intelligence, 2019).

4. Empirical Research and Results

The basis for this paper stems from the search for an answer to the following research question:

Which ways are users aware of unethical artificial intelligence and it’s mitigation measures?

Answering this question is important because the trend is to integrate more and more algorithmic systems into our daily lives. In addition to the positive aspects, AI systems also pose certain risks that users should be aware of (Ada Lovelace Institute, 2020; Raji et al., 2020). However, a majority of users are unaware of this impact on their lives (Gran et al., 2021; Shin et al., 2022). Therefore, this study examines whether the results of the studies apply to this survey population.

The first hypothesis tests the relationship between self-reported awareness and users' self-efficacy of algorithmic platforms. The following four hypotheses analyze the relationship between the FEAT factors and users' self-efficacy. Knowledge about the functionality and inner workings of algorithms can improve users' confidence in using them (Kizilcec, 2016). This theory is tested by hypothesis one (H1).

Self-reported awareness of algorithms positively influences users' self-efficacy of algorithmic platforms. (H1)

Hypothesis two (H2) is tested because algorithmic fairness is connected to users' attitudes, specifically to their confidence (Alter, 2021).

Perceived fairness positively influences users' self-efficacy of algorithmic platforms. (H2)

Studies suggest that if the users are able to understand how their data is collected and processed, hence understand how recommendations are generated, they are more able to adopt to the system (Rai, 2020). This relation will be tested by hypothesis three (H3).

Perceived explainability positively influences users' self-efficacy of algorithmic platforms. (H3)

Through clearly defined roles, responsibilities and liabilities, users are able to establish positive self-assurance (Diakopoulos, 2015). Therefore, it is hypothesized that perceived accountability has an influence on the users' self-efficacy (H4).

Perceived accountability positively influences the users' self-efficacy of algorithmic platforms. (H4)

When processes are evident and transparent, consumers are more likely to interact with the outcomes in a more trustworthy and engaging manner (Gran et al., 2021; Renjith et al., 2020). This is why hypothesis five (H5) suggests that through transparent systems the user gains assurance, hence an impression of efficacy.

Perceived transparency positively influences the self-efficacy of algorithmic platforms. (H5)

Building trust with a user requires a certain level of confidence, competence, and transparency. Because when the user knows how to use algorithms, trust develops (Hu et al., 2021; Rader, 2017). This is tested through hypothesis six (H6).

Self-efficacy positively influences users' trust in algorithmic platforms. (H6)

Trust and confidence in the algorithmic platform can enhance the awareness and knowledge about mitigation measures, which ensure that algorithmic platforms are more secure and less biased (Raji et al., 2020; Shin et al., 2022). This relation will be tested by hypothesis seven (H7).

Trust positively influences the awareness of mitigation measures. (H7)

As education plays a crucial role in developing algorithmic awareness, experience and interest in the field of information technology can help increase algorithmic awareness and awareness

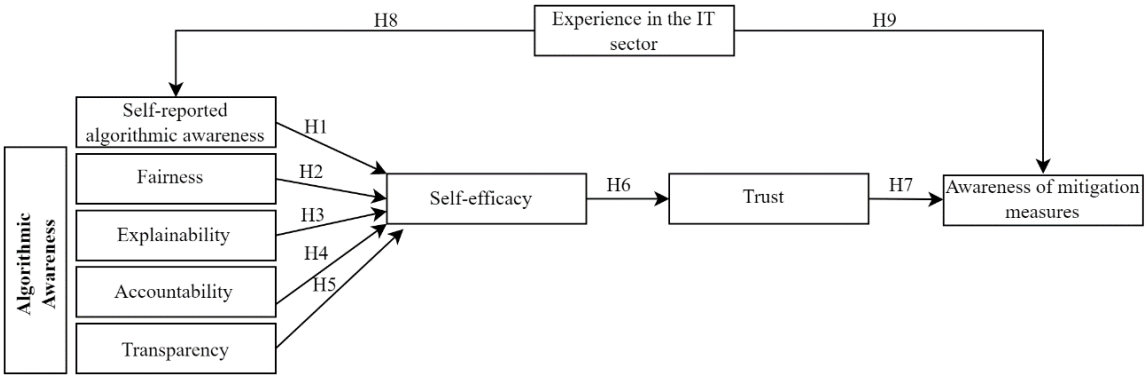
of mitigation measures for unethical AI (Gran et al., 2021). This is tested by hypothesis eight (H8) and nine (H9).

Experience in the IT Sector positively influences the self-reported algorithmic awareness. (H8)

Experience in the IT Sector positively influences the Awareness of Mitigation Measures. (H9)

Figure 1 shows the relationship of the hypotheses.

Figure 1. Hypothesis Path Tree



Source: Own research

The main difference to the study of Shin et al. (2022) is that this research tries to connect the algorithmic awareness with the awareness on mitigation measures. Shin et al. (2022) focused additionally on the privacy aspect and users' self-disclosure, which are variables that have been excluded from this study, due to time and resource restrictions.

In this study, all measures were derived from the literature where they were validated. The measures were adapted to the current study and only additional explanations were added where the pretest showed that this was necessary to improve the comprehensibility of the questions.

In order to test the hypothesis with a representative sample corresponding to the use of algorithmic platforms, the online format was unavoidable. It ensures that participants are more likely to use algorithmic platforms and are familiar with them. The questionnaire was created using the online tool SoSci Survey.

The first part of the survey examines the level of AA. The variables tested are self-reported algorithmic awareness, explainability, accountability, fairness and transparency. The following questions are about the user's self-efficacy and trust in the algorithmic platform, as well as mitigation measures. The only question with a 5-point Likert scale is the variable self-reported algorithmic awareness; the rest of the questions have a 7-point Likert scale (see Appendix, ***Table 6***). The final section collects demographic data such as age, gender, current location, highest level of education completed, and years of experience in IT.

The survey was online for 18 days and collected a total of 381 responses. However, 291 of these responses are valid for analysis after cleaning the dataset.

The distribution of the sample is that 52% were of the female gender, with an average age of 29 years. 48% were of the male gender, with an average age of 29 years. In order to adequately capture gender diversity, the options "other" and "prefer not to say" were included. These were clicked on 0.7% of the time, with an average age of 31 years. As the questionnaire was open to everyone, the age distribution is quite diverse. However, the majority of responses came from people between the ages of 18 and 35 (81%), which is not surprising given the social network

of the principal investigator. In order to obtain a heterogeneous sample, the survey was available in German and English and could be answered worldwide. As a result, the current residence of the participants is spread over 33 countries. 75% were from the following countries: Austria (53%), Germany (9%), France (8%), and the United States of America (6%). Respondents were asked about their current location, as their current environment and location can change their awareness of algorithms (Gran et al., 2021).

Education is an important factor in algorithm awareness. The study by Gran et al. (2021) found that the higher the level of education, the more critical of algorithmic platforms the respondents in their study were. In this study, 78% had a university degree. In addition, experience and interest in information technology (IT) may influence attitudes. In this study, 54% have some experience in IT. Detailed demographics are shown in ***Table 7*** in the Appendix.

A confirmatory factor analysis (CFA) was conducted to assess the reliability and validity of the variables. The model was tested for its indicator reliability, concurrent reliability, convergent validity, and discriminant validity (Hair et al., 2017b).

To test the reliability of reflective indicators, the factor loadings are inspected. According to Fronell & Larcker (1981) the factor loading of each indicator needs to be equal or above 0.7, so that the shared variance amidst the construct and its indicators is greater than the variance of the error term. ***Table 1*** shows the factor loadings of the data set. The results of the first loading calculation are that Q11_AC3 and Q5_T2 have a negative loading and therefore need to be excluded from the analysis. The loading of Q13_F2 was only 0.144, so it did not meet the required threshold and must be excluded (Hair et al., 2017b). Within the second loading calculation, all values are higher than 0.7.

Internal consistency reliability can be measured by Cronbach's Alpha (CA) and Composite Reliability (CR). The difference between these two measures is that CA assumes that all indicators are equally reliable, whereas CR takes into account these dissimilarities in item reliability. PLS-SEM prioritizes indicators based on their reliability. Therefore, CR values are more appropriate for PLS-SEM. The CR should exceed a threshold of 0.7 (Hair et al., 2017b). This is the case for all factors, which can be seen in Table 1, and the internal consistency reliability can be approved. Convergent validity is checked by the average variance extracted (AVE). Based on the literature, the threshold of AVE values must be 0.5 or higher (Fornell & Larcker, 1981). The constructs in the study meet this criterion, as can be seen in Table 1.

Table 1. Measurement Model Assessment

Constructs	Factors	Loadings Before Cleaning	Loadings After Cleaning	CR	AVE
Self-reported algorithmic awareness	Q3_A1	1.00	1.00	1.00	1.00
Fairness	Q4_F1	0.81	0.83	0.87	0.69
	Q13_F2	0.14	**		
	Q14_F3	0.85	0.86		
	Q8_F4	0.80	0.80		
Explainability	Q6_E1	0.90	0.90	0.88	0.71
	Q9_E2	0.85	0.85		
	Q7_E3	0.77	0.77		
Accountability	Q15_AC1	0.58	0.73	0.74	0.59
	Q10_AC2	0.57	0.80		
	Q11_AC3	-0.76	**		

Constructs	Factors	Loadings Before Cleaning	Loadings After Cleaning	CR	AVE
Transparency	Q12_T1	0.78	0.93	0.84	0.72
	Q5_T2	-0.22	**		
	Q16_T3	0.64	0.77		
Experience in IT sector	Q48_it	1.00	1.00	1.00	1.00
Self-efficacy	Q21_S1	0.78	0.78	0.85	0.65
	Q18_S2	0.78	0.78		
	Q23_S3	0.85	0.85		
Trust	Q17_TR1	0.85	0.85	0.90	0.76
	Q22_TR2	0.87	0.87		
	Q19_TR3	0.89	0.89		
Awareness of mitigation measures	Q20_M1	1.00	1.00	1.00	1.00
	Q24_M2	*			
	Q25_M3	*			

*Q24_M2 was a text questions and Q25_M3 was a multiple-choice question, therefore they will be analyzed separately.
** As the loadings of Q11_AC3 and Q5_T2 are negative, they will be excluded from the analysis. Additionally, Q13_F2 was excluded as its loading was under the threshold of 0.7 (Hair, Hult, et al., 2017).
Source: Own results

To determine discriminant validity, the Fornell-Larcker criterion was examined. To test the validity of the criterion, the square root of the construct AVE value should be greater than the correlation with other constructs. This is shown in ***Table 2***. Since the values meet this rule, this criterion can be accepted.

Table 2. Fornell-Larcker Criterion

	1	2	3	4	5	6	7	8	9
Accountability	0.77								
Self-reported algorithmic awareness	0.17	1.00							
Explainability	-0.18	-0.02	0.84						
Fairness	-0.03	-0.24	0.34	0.71					
IT experience	0.02	0.23	0.07	-0.02	1.00				
Awareness of mitigation measures	0.27	0.05	0.05	0.06	-0.03	1.00			
Self-efficacy	0.02	0.12	0.12	0.14	0.01	0.32	0.80		
Transparency	0.48	0.09	0.09	-0.12	-0.11	0.11	-0.11	0.85	
Trust	-0.09	-0.05	-0.05	0.39	-0.01	0.22	0.46	-0.24	0.87

Source: Own results

In this paper, the reliability and validity of the measurement model are considered satisfactory, as the loading of each item is higher than 0.7. The CR is greater than 0.7. The AVE value is greater than 0.5 and the Fornell-Lacker criterion is met.

The goodness of model fit in Smart PLS is recommended for use with CB-SEM models. Since this model is calculated based on PLS-SEM, the reliability and validity measures mentioned above are sufficient to determine the goodness of model fit (Hair et al., 2017a).

Before analyzing the structural model, a bootstrapping technique was applied. It was calculated with 5000 subsamples of the 291 samples from the study (Hair et al., 2011). Bootstrap confidence intervals are obtained using a two-tailed test with a 5% significance level (Hair et al., 2017b).

By checking the variance inflation factors (VIF), it was determined that there are no multicollinearity problems. Next, the path coefficients are analyzed to see if there are significant relationships. The values should be between -1 and +1, so higher values indicate a stronger positive relationship between the variables. The path coefficients were calculated using a two-tailed test with a significance level of 5%. Therefore, the t-values should be above 1.97 and the p-values below 0.05, indicating a significant relationship. The values are shown in ***Table 3***.

Table 3. Path Coefficients

	Sample Mean (M)	Standard Deviation (STD)	T-Statistics	p-values	Results
Accountability – self-efficacy	0.08	0.10	1.04	0.300	not significant
Self-reported algorithmic awareness – self-efficacy	0.16	0.05	**2.81**	**0.005**	**significant**
Explainability – self-efficacy	0.19	0.06	**3.20**	**0.001**	**significant**
Fairness – self-efficacy	0.11	0.06	1.61	0.107	not significant
IT experience – self-reported algorithmic awareness	0.23	0.05	**4.63**	**0.000**	**significant**
IT experience – awareness of mitigation measures	-0.04	0.06	0.64	0.524	not significant
Self-Efficacy - awareness of mitigation measures	0.29	0.07	**3.75**	**0.000**	**significant**
Self-Efficacy - trust	0.46	0.05	**8.39**	**0.000**	**significant**
Transparency – self-efficacy	-0.10	0.06	1.72	0.086	not significant
Trust – awareness of mitigation measures	0.09	006	1.38	0.167	not significant

Source: Own results

A potential mediation effect can be found in the research model. Therefore, this effect was tested according to the guidelines of Hair et al. (2017b). The effect of trust mediating the relationship between self-efficacy and awareness of mitigation measures is examined in Smart PLS. The indirect effect between these constructs is the product of the standardized paths.

The very common Sobel test is not applicable to PLS-SEM analysis. Instead, the bootstrapping technique is used to analyze the mediation effect (Hair et al., 2017b). Bootstrapping is used to test the total and indirect effects of self-efficacy on awareness of mitigation measures. The results are shown in ***Table 4***. The direct effect value is the path coefficient of the direct relationship between self-efficacy and awareness of mitigation measures. The indirect effect score is the product of the path coefficient from self-efficacy to trust and from trust to awareness of mitigation measures. As Table 4 shows, the direct effect is significant while the indirect effect is not. This implies that there is no direct mediation. Therefore, there is no mediation because there is a significant direct effect of self-efficacy on awareness of mitigation measures.

Table 4. Indirect and direct effects of self-efficacy on awareness of mitigation measures

	Effect value	Sample Mean (M)	Standard Deviation (STD)	T-statistics	p-values
Direct effect					
Self-efficacy – awareness of mitigation measures	0.28	0.29	0.08	3.72	0.00
Indirect effect					
Self-efficacy – awareness of mitigation measures	0.04	0.04	0.03	1.31	0.19

Source: Own results

R2 tests the predictive power of the research model (Hair et al., 2011). The results are summarized in ***Table 5***. It shows that the variables have very little or no predictive power on the constructs of endogenous nature. Only the variable trust comes close to the value of a weak classification. In addition, the Stone-Geisser Q2 value can be used to evaluate the predictive relevance of the model (Geisser, 1974). The R2 value assesses the in-sample predictive power, while the Q2 value measures the out-of-sample predictive power of the model (Hair et al., 2017b). Within this research data, blindfolding was calculated with a D of 7. As shown in Table 5, all values for the endogenous constructs are above zero. Therefore, the path model is able to estimate the originally collected values (Hair et al., 2017b).

Table 5. Prediction Power and Predictive Relevance - R2 and Q2 values

	R^2	t-value	p-value	Q^2
Self-reported algorithmic awareness	0.05	2.81	0.05	0.05
Awareness of mitigation measures	0.11	0.64	0.06	0.08
Self-efficacy	0.09	3.75	0.07	0.04
Trust	0.21	8.39	0.05	0.16

Source: Own results

In this paper, the structural model assessment included the following results:

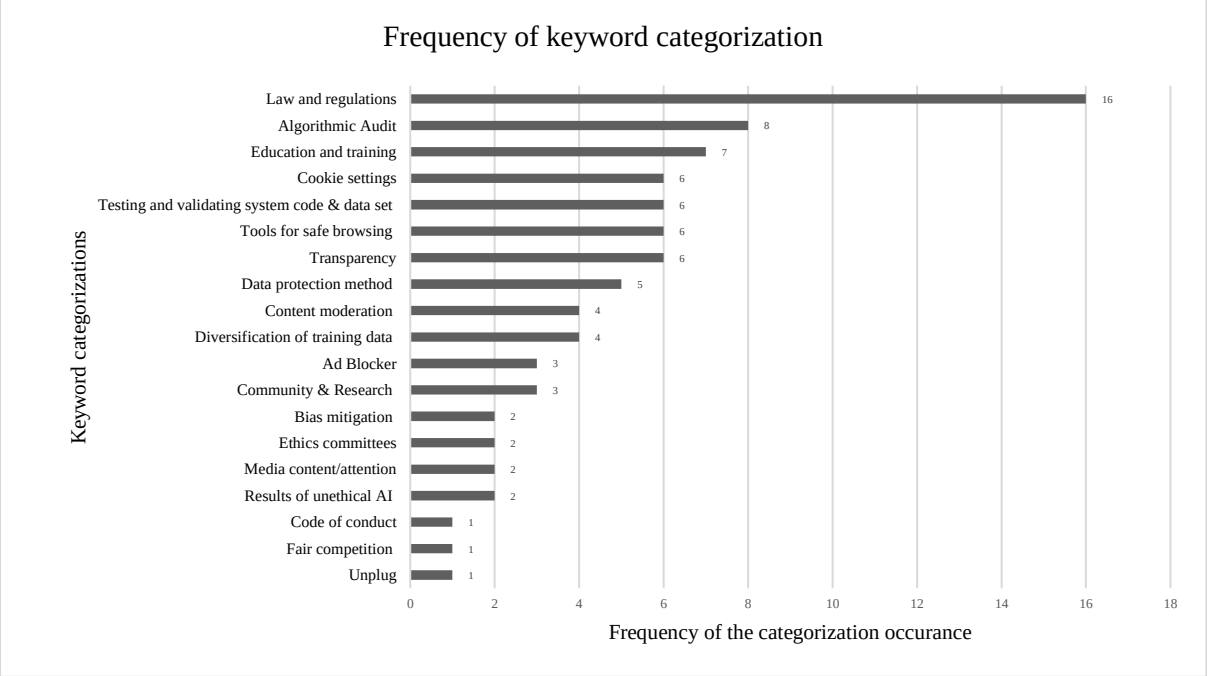
No multicollinearity problems are included in the data, as the VIF values are underneath the threshold of 3.3. The path coefficients have been analyzed for each hypothesis. Hence, the following hypotheses have been accepted: H1, H3, H6, H8. Mediation effect of Trust on Self-Efficacy and Awareness of Mitigation Measures has been tested and resulted in a direct-only non-mediation. The coefficient of determination (R^2) has been examined for each variable, evaluating that the variables only have little to no prediction power. With the predictive accuracy (Q^2), it has been analyzed that the path model is capable of estimating the original collected values. Furthermore, the effect sizes have been investigated with Cohen f^2 and q^2. The first resulted in small to medium effects of some constructs and the latter showed a very low predictive relevance on the endogenous constructs.

Questions Q24_M2 and Q25_M3 will be analyzed in the following using the tool RStudio and the programming language R. Question Q24_M2 asked participants to enter mitigation measures against unethical artificial intelligence that came to mind. The question was optional and was answered 55 times. By answering the question, participants showed their familiarity with and understanding of the topic. In addition, it can be analyzed if other mitigation measures that do not appear in the literature play a role in people's lives.

In order to obtain results from this open question, certain methods of text analysis were applied. The word bubble in ***Figure 2*** visualizes the most frequent words. After cleaning, the text can be categorized. Since the dataset is quite small, with only 56 entries consisting mainly of keywords, a manual rule-based approach is the most suitable for categorization. To do this, certain rules are defined, for example, if the word "GDPR" appears, the label "laws and regulations" is assigned (Zhai & Massung, 2016).

The next question comes right after the open question. Q25_M3 was added to the questionnaire to better analyze whether participants understood what mitigation measures for unethical AI are. The multiple-choice question contained 18 options, two of which were made up and should not have been selected as mitigation measures. These two options were named "Ignore the unethical practices" and "Spread fake news".

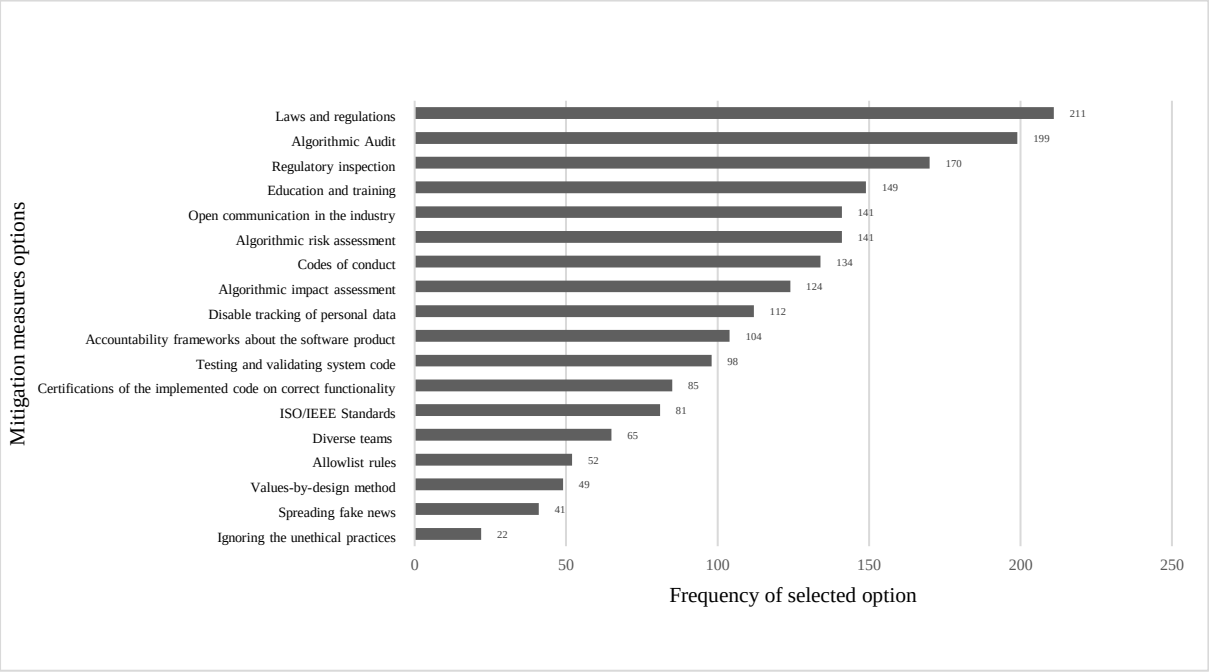
Figure 2. Frequency of Keyword Categorization of Question Q24_M2



Source: Own results

On average, six responses were selected. ***Figure 3*** shows the frequency count of each possible response option. The most selected answers were "Laws and regulations" with 211 frequencies, "Algorithmic audit" with 199 frequencies, and "Regulatory inspection" with 170 frequencies. The least selected were the two invented mitigation measures and "Values-by-design method".

Figure 3. Frequency of selected options on mitigation measures against unethical AI



Source: Own results

The result shows that according to the participants, "Laws and regulations" is the most logical mitigation measure against unethical AI. The four terms "algorithmic audit", "regulatory inspections", "algorithmic risk assessments", and "algorithmic impact assessments" were also selected quite often. In between are the options regarding "education and training" and "open communication in the industry". Other options related to tracking, testing, frameworks, certifications and standards were not selected as frequently.

5. Discussion

This paper contributes to the development of a model approach for AA processing in the context of algorithms. This model explains how interaction with algorithms involves certain AA processes that provide a basis for generating heuristics of user motivation and triggers algorithm adoption behavior (Shin et al., 2022). Thus, the model shows which attributes contribute to a particular impression of effectiveness and how general algorithm awareness affects awareness of mitigation measures.

The results of this paper add relevant insights to the current literature on the relationship between awareness, trust, and understanding of mitigation measures. The study explored the following main points: self-reported awareness of algorithmic awareness, awareness of unethical AI, and its mitigation measures. The questionnaire asked participants to rate their algorithmic awareness, which can be categorized as a subjective analysis of general awareness of algorithms (Gran et al., 2021). The results have shown that there is a significant relationship between experience in the IT sector and the self-reported algorithmic awareness (H8). Therefore, it can be said that a certain degree of interest, knowledge and experience in the IT sector contributes to the self-reported algorithmic awareness. The connection between self-reported algorithmic awareness and self-efficacy has been accepted (H1). This implies that people, who believe they are algorithmic aware, do have a better impression of their skillfulness to perform tasks within algorithmic platforms. It also affects their ability to learn and acquire knowledge about the domain. Research has shown that more information increases trust in the algorithmic platform (Shin et al., 2022). This result can also be accepted by this paper, as the relationship between self-efficacy and trust is significant (H6). Based on this results, trust plays a crucial role in connection with AA and Mitigation Measures. If the user trusts the algorithmic platform, it is less likely that the user has concerns of being unethically treated by it. Hence, the role of mitigation measures of unethical AI is diminished, which can be seen in the non-significant relationship of trust with awareness of mitigation measures (H7). This results in the only-direct non-mediating effect of trust on the connection of self-efficacy to awareness of mitigation measures. What implies that the knowledge and skillfulness of the user influences their awareness of mitigation measures. The findings of Shin et al. (2022) described that the FEAT factors should be considered in algorithm design and operation. However, within this study, it resulted in a non-significant relationship of the factors fairness, accountability and transparency in connection with self-efficacy (H2, H4, H5). These three factors mainly investigated how the agreement of the participants in connection with unethical AI is. Thus, these three factors do not influence the level of performance of the participants in the operation of algorithmic platforms, whereas the factor explainability has a significant influence on self-efficacy (H3). Based on this positive relationship, it can be recommended that algorithmic platform providers focus more on explainable AI (XAI). As already explained, self-efficacy does have a direct significant relationship with awareness of mitigation measures; hence the skillfulness of people has a certain influence on their awareness of mitigation measures.

However, the relationship between experience in the IT sector and awareness of mitigation measures is not significant (H9).

In summary, it means that experience in the IT sector does have an influence on the self-reported algorithmic awareness, which has a positive relation with self-efficacy. The mediator effect of self-efficacy to awareness of mitigation measures is significant; hence, the ability of users to perform certain tasks and acquire knowledge does influence their awareness of mitigation measures. However, trust, which is also significant with self-efficacy, has no relation to the awareness of mitigation measures. Which in turn means, that if users put trust in the algorithmic platform, they will be hindering themselves to gain knowledge or skills about mitigation measures against unethical AI. Interestingly, experience in the IT sector does not imply to have influence on awareness or knowledge of mitigation measures.

The open text question, which was optional, was designed to analyze participants' awareness of mitigation measures and whether their ideas were consistent with the literature. A total of 55 participants provided valuable responses, resulting in 85 keywords. Of these keywords, sixteen were related to "laws and regulations", eight were related to "algorithmic audit" practices, and seven were related to "education and training" (Figure 2). Based on the high number of keywords related to "cookie settings", "safe browsing tools" or "ad blockers", it can be said that these participants are more prone to the practical tools that they can use on a daily basis. Other frequently occurring factors are "testing and validating system code & dataset" and "diversifying training data". Terms not found in the literature include "content moderation," which was mentioned four times, "media content/attention," which occurred twice, "ethics committees," which was mentioned twice, and "fair competition," which was mentioned once. On the other hand, participants did not mention frameworks, certifications or standards. Keywords included the importance of diverse data sets, but not the importance of diverse teams (mentioned in the literature). It should be noted, however, that the sample size for the open-ended question is too small to draw any reliable conclusions.

There is some interpretable overlap between the responses to the open text question and the multiple-choice mitigation question. In both questions, the three most popular options are "law and regulation", various types of "algorithmic testing", and "education and training". Therefore, it can be understood that the most resonant mitigation measures are located in these three areas. The research question of this paper investigates approaches to create algorithmic awareness among users. Based on the empirical study, the following conclusions can be drawn Experience in the IT sector has an influence on algorithmic awareness. Three variables of the FEAT factors of the AA model of Shin et al. (2022) have no influence on self-efficacy; therefore, explainability has a positive influence on self-efficacy and thus on awareness of mitigation measures. Trust in algorithmic platforms has no significant relationship with awareness of mitigation measures. The respondents of the study consider the three areas of law and regulation, algorithmic audit, and education and training to be the most accessible to them. It can be concluded that these three areas should be strengthened in order to create ethical, trustworthy and transparent algorithmic platforms and to increase the awareness of algorithmic platform users.

6. Conclusion

Our societies depend on rapidly produced artificial intelligence systems that have rarely been examined to see if these systems do not amplify, accentuate, and systematize human biases

(Raji et al., 2020). As research has shown, this is a problem because the public is unaware of the impact of algorithmic decisions on their lives (Gran et al., 2021; Shin et al., 2022). This is why this paper investigated the level of self-reported algorithmic awareness and their consciousness of unethical artificial intelligence and its mitigation measures. This paper improves the understanding of the relation of self-reported algorithmic awareness, the AA model from Shin et al. (2022), experience in the IT sector and the awareness of mitigation measures of unethical AI. It can be said that experience in the IT sector does positively influence the self-reported algorithmic awareness, which influences the users' self-efficacy. Hence, the awareness of mitigation measures is affected as the user is able to navigate through the algorithmic platform and gain knowledge. However, when the user is trusting the algorithmic platform the building of algorithmic awareness is disturbed. Thus, trusting the algorithmic platform will hinder the user to acquire knowledge or skills about mitigation measures against unethical AI. The analysis showed that the participants of the study are more aware of mitigation measures, which are in the areas *law and regulations*, *algorithmic audit* and *education and training*. Hence, these three fields should be elaborated on, in order to establish ethical, trustful and transparent algorithmic platforms and to deepen the understanding and awareness of the users of algorithmic platforms. In conclusion, this paper contributed valuable empirical findings for researcher and practitioners in the ethical AI field.

References:

Ada Lovelace Institute. (2020). *Examining the Black Box*. https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf

Alter, S. (2021). Understanding artificial intelligence in the context of usage: Contributions and smartness of algorithmic capabilities in work systems. *International Journal of Information Management*, *May*, 102392. https://doi.org/10.1016/j.ijinfomgt.2021.102392

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies_Ashok.pdf. *International Journal of Information Management*, *62*(Feburary). https://doi.org/10.1016/j.ijinfomgt.2021.102433

Bazerman, M. H., & Moore, D. A. (2012). *Judgment in managerial decision making (3rd ed.)* (8th editio). Wiley.

Bereby-Meyer, Y., & Grosskopf, B. (2008). Overcoming the winner's curse: An adaptive learning perspective. *Journal of Behavioral Decision Making*, *21*(1), 15–27. https://doi.org/10.1002/BDM.566

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*.

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Diakopoulos, N. (2015). Accountability in algorithmic decision-making. *Queue*, *13*(9), 126–149. https://doi.org/10.1145/2857274.2886105

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First I "like" it, then I hide it: Folk theories of social feeds. *Conference on Human Factors in Computing Systems - Proceedings*, 2371–2382. https://doi.org/10.1145/2858036.2858494

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news

feeds. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*, 153–162. https://doi.org/10.1145/2702123.2702556

Eticas Consulting. (2021). *Guide to Algorithmic Auditing*. *January*.

European Commision's High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. In *European Commission*.

Fast, N. J., & Jago, A. S. (2020). Privacy matters… or does It? Algorithms, rationalization, and the erosion of concern for privacy. *Current Opinion in Psychology*, *31*, 44–48. https://doi.org/10.1016/j.copsyc.2019.07.011

Fazal-E-Hasan, S. M., Ahmadi, H., Mortimer, G., Lings, I., Kelly, L., & Kim, H. (Jay). (2020). Online Repurchasing: The Role of Information Disclosure, Hope, and Goal Attainment. *Journal of Consumer Affairs*, *54*(1), 198–226. https://doi.org/10.1111/joca.12263

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39. https://doi.org/10.2307/3151312

Fryer, L. K., Thompson, A., Nakao, K., Howarth, M., & Gallacher, A. (2020). Supporting self-efficacy beliefs and interest as educational inputs and outcomes: Framing AI and Human partnered task experiences. *Learning and Individual Differences*, *80*(February). https://doi.org/10.1016/j.lindif.2020.101850

Geisser, S. (1974). Biometrika Trust A Predictive Approach to the Random Effect Model Author (s): Seymour Geisser Published by: Oxford University Press on behalf of Biometrika Trust Stable URL : http://www.jstor.org/stable/2334290. *Biometrika Trust*, *61*(1), 101–107.

Gran, Booth, P., & Bucher, T. (2021). To be or not to be algorithm aware: A question of a new digital divide? *Information, Communication & Society*, *24*(12), 1779–1796. https://doi.org/10.1080/1369118X.2020.1736124

Hair, J., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L. (2017a). An updated and expanded assessment of PLS-SEM in information systems research. *Industrial Management and Data Systems*, *117*(3), 442–458. https://doi.org/10.1108/IMDS-04-2016-0130

Hair, J., Hult, T., Ringle, C., & Sarstedt, M. (2017b). A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)—Joseph F. Hair, Jr., G. Tomas M. Hult, Christian Ringle, Marko Sarstedt. In *Sage*.

Hair, J., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Hamilton, K., Sandvig, C., Karahalios, K., & Eslami, M. (2014). A path to understanding the effects of algorithm awareness. *Conference on Human Factors in Computing Systems - Proceedings*, 631–640. https://doi.org/10.1145/2559206.2578883

Hargittai, E., Gruber, J., Djukaric, T., Fuchs, J., & Brombach, L. (2020). Black box measures? How to study people's algorithm skills. *Information Communication and Society*, *23*(5), 764–775. https://doi.org/10.1080/1369118X.2020.1713846

Hu, Q., Lu, Y., Pan, Z., Gong, Y., & Yang, Z. (2021). Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants. *International Journal of Information Management*, *56*(September 2020), 102250. https://doi.org/10.1016/j.ijinfomgt.2020.102250

Kizilcec, R. F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395. https://doi.org/10.1145/2858036.2858402

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law and Security Review*, *34*(4), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017

Micocci M, Borsci S, Thakerar V, Walne S, Manshadi Y, Edridge F, Mullarkey D, Buckle P, & Hanna, G. (2021). GPs Trust Artificial Intelligence Insights and What Could This Mean for Patient Care? A Case Study on GPs Skin Cancer Diagnosis in the UK. . *Preprints 2021, 2021050005*, *May*.

Park, S., & Humphry, J. (2019). Exclusion by design: Intersections of social, digital and data exclusion. *Information Communication and Society*, *22*(7), 934–953. https://doi.org/10.1080/1369118X.2019.1606266

Powers, E. (2017). My news feed is filtered? Awareness of news personlization among college students. *Digital Journalism*. https://doi.org/10.1080/21670811.2017.1286943

Rader, E. (2017). Examining user surprise as a symptom of algorithmic filtering. *International Journal of Human Computer Studies*, *98*(December 2015), 72–88. https://doi.org/10.1016/j.ijhcs.2016.10.005

Rader, E., & Gray, R. (2015). Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 173–182. https://doi.org/10.1145/2702123.2702174

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*. https://doi.org/10.1145/3351095.3372873

Renjith, S., Sreekumar, A., & Jathavedan, M. (2020). An extensive study on the evolution of context-aware personalized travel recommender systems. *Information Processing and Management*, *57*(1), 102078. https://doi.org/10.1016/j.ipm.2019.102078

Rovatsos, M., Mittelstadt, B., & Koene, A. (2019). Landscape Summary: Bias in Algorithmic. *Centre for Data Ethics and Innovation*.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*. 1–23.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. In *International Journal of Human Computer Studies* (Vol. 146). https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D., Kee, K. F., & Shin, E. Y. (2022). Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? *International Journal of Information Management*, *65*, 102494. https://doi.org/10.1016/j.ijinfomgt.2022.102494

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance | Elsevier Enhanced Reader. *Computers in Human Behavior*, *98*(September), 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Supreme Audit Institutions of Finland Germany the Netherlands Norway and UK. (2020, November 24). *Auditing machine learning algorithms*. https://auditingalgorithms.net/index.html

Swart, J. (2021). Experiencing Algorithms: How Young People Understand, Feel About, and Engage With Algorithmic News Selection on Social Media. *Social Media and Society*, *7*(2). https://doi.org/10.1177/20563051211008828

Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). *Telematics and Informatics*, *62*(December 2020), 101607. https://doi.org/10.1016/j.tele.2021.101607

Zhai, C., & Massung, S. (2016). *Text Data Management and Analysis*.

Zhang, Y., Vera Liao, Q., & Bellamy, R. K. E. (2020). Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852

Appendix

Table 6. Overview of survey questions

Question Code	Question	Variable	Type
Q1	Welcome Text		Text
Q2	Explanation of Algorithm Platforms		Text
Q3_A1	What kind of awareness do you have that algorithms are used to present recommendations, advertisements and other content on the internet?	Self-reported Algorithmic Awareness	5-point Likert Scale (awareness)
Q4_F1	An algorithmic platform does not discriminate against people.	Fairness (Non-discrimination)	7-point Likert Scale (agreement)
Q5_T2	Any results generated by an algorithmic system should be interpretable to the users affected by those outputs.	Transparency (Interpretability)	7-point Likert Scale (agreement)
Q6_E1	I find algorithmic platforms to be comprehensible.	Explainability	7-point Likert Scale (agreement)
Q7_E3	I can understand and make sense of the internal workings of personalization.	Explainability	7-point Likert Scale (agreement)
Q8_F4	An algorithmic platform does not show favoritism.	Fairness (Non-discrimination)	7-point Likert Scale (agreement)
Q9_E2	Algorithmic platforms are understandable.	Explainability	7-point Likert Scale (agreement)
Q10_AC2	The platforms should be designed to enable third parties to audit and review the behavior of an algorithm.	Accountability (Auditability)	7-point Likert Scale (agreement)
Q11_AC3	The platform operators should have the autonomy to change the logic in their entire configuration using only simple manipulations (for example by changing the sorting of information).	Accountability (Controllability)	7-point Likert Scale (agreement)
Q12_T1	The assessment and the criteria of used algorithms should be publicly open and interpretable by the users.	Transparency (Understandability)	7-point Likert Scale (agreement)
Q13_F2	The source of data throughout an algorithmic process and its data analysis should be accurate and correct.	Fairness (Accuracy)	7-point Likert Scale (agreement)
Q14_F3	An algorithmic platform is impartial.	Fairness (Due process)	7-point Likert Scale (agreement)
Q15_AC1	It should be required, that the person in charge of the algorithmic platform requires (for example the operator of the platform) to be made accountable for its adverse individual or societal effects (for example miscalculations, discriminatory categorization etc.) in a timely manner.	Accountability (Responsibility)	7-point Likert Scale (agreement)
Q16_T3	Algorithmic platforms should inform about their internal procedures.	Transparency (Observability)	7-point Likert Scale (agreement)
Q17_TR1	I trust the recommendations by algorithm-driven platforms.	Trust	7-point Likert Scale (agreement)
Q18_S2	I am certain that I can work effectively on different tasks (for example online shopping or information research) in my interactions with algorithmic platforms.	Self-Efficacy	7-point Likert Scale (agreement)

Question Code	**Question**	**Variable**	**Type**
Q19_TR3	The personalized results from the algorithmic platform are trustworthy.	Trust	7-point Likert Scale (agreement)
Q20_M1	Audits, where the algorithm is inspected, help to discover any unethical behavior.	Awareness of Mitigation Measures (MM)	7-point Likert Scale (agreement)
Q21_S1	On algorithmic platforms I can carry out the planned tasks (for example a search request) most of the times.	Self-Efficacy	7-point Likert Scale (agreement)
Q22_TR2	Recommended results via algorithmic processes are credible.	Trust	7-point Likert Scale (agreement)
Q23_S3	When facing an urgent situation (for example immediate online search), I am positive that I can accomplish it through algorithmic platforms.	Self-Efficacy	7-point Likert Scale (agreement)
Q24_M2	Which mitigation measures against unethical artificial intelligence do you know? (optional)	Awareness of Mitigation Measures (MM)	Open text
Q25_M3	Which of the following would you identify as mitigation measures for unethical algorithmic platforms?	Awareness of Mitigation Measures (MM)	Multiple Choice
Q26_M3	Algorithmic Audit	MM option	
Q27_M3	Regulatory Inspection	MM option	
Q28_M3	Algorithmic Risk Assessment	MM option	
Q29_M3	Algorithmic Impact Assessment	MM option	
Q30_M3	Laws and regulations	MM option	
Q31_M3	Ignoring the unethical practices	MM option	
Q32_M3	Allowlist rules	MM option	
Q33_M3	Values-by-design method	MM option	
Q34_M3	Testing and validating system code	MM option	
Q35_M3	Disable tracking of personal data	MM option	
Q36_M3	Codes of conduct	MM option	
Q37_M3	Spreading fake news	MM option	
Q38_M3	ISO/IEEE Standards	MM option	
Q39_M3	Certifications of the implemented code on correct functionality	MM option	
Q40_M3	Accountability frameworks about the software product	MM option	
Q41_M3	Open communication in the industry	MM option	
Q42_M3	Education and training	MM option	
Q43_M3	Diverse teams	MM option	
Q44_age	How old are you?	Age	Text
Q45_gender	Which gender do you identify yourself with?	Gender	Single-choice
Q46_country	Which is the country, you're currently living?	Country	Text
Q47_education	What is your highest completed education?	Education	Single-choice
Q48_it	Do you have experience in the information technology (IT) field?	It	Single-choice

Table 7. Data description of valid responses

Category	Respondents (n)	Percentage (%)
Gender		
Female	150	51.55 %
Male	139	47.76 %
Diverse	0	-
Rather not say	2	0.69 %
Age		
18-25	146	50.18 %
26-35	90	30.93 %
36-45	38	13.06 %
46-55	12	4.12 %
55-66	5	1.71 %
Current residence		
Austria	153	52.57 %
Germany	25	8.60 %
France	23	7.90 %
United States of America	18	6.19 %
Slovakia	6	2.06 %
Lebanon	6	2.06 %
Switzerland	5	1.72 %
Netherlands	5	1.72 %
Other countries	50	17.18 %
Highest completed occupation		
Apprenticeship	1	0.35 %
Elementary school	3	1.03 %
A-levels	49	16.84 %
Bachelor’s degree	140	48.10 %
Diploma/Master’s degree	87	29.90 %
PhD	11	3.78 %
None of the above	0	-
Experience in information technology (IT) field		
No	135	46.39 %
Less than two years	57	19.59 %
More than two years	99	34.02 %

Table 8. Measurement of variables

Independent variables	
Fairness:	The questions around the variable fairness tested whether the respondents think that algorithmic platforms treated them fairly and consistently with the laws and principles (Shin & Park, 2019). The variable consists out of four questions, with the codes Q4_F1, Q13_F2, Q14_F3 and Q8_F4. The first and second questions aimed to explore the non-discriminatory character of algorithmic platforms. For a better understandability, the original question was split in two, as it asked for two different aspects of nondiscrimination. The third one tested the accuracy of the source data used and the last question was aimed at analyzing the due process.
Explainability:	With the questions about explainability, it was aimed to test if users' decision making is eroded, when using the algorithmic platforms (Rai, 2020; Shin, 2021). The variable includes three questions with the codes Q6_E1, Q7_E3 and Q9_E2.
Accountability:	This variable tests if designers and/or providers of the algorithmic platform should be held responsible for the results they caused by providing the service (Diakopoulos, 2015). It is explored through three questions, Q15_AC1, Q10_AC2 and Q11_AC3. The first question specifically tests the responsibility. The second question examines the auditability and the third one the controllability.
Transparency:	This term means that algorithm-generated decisions should be unclosed, provable and/or apparent to the users, because they are the ones who are consuming, adopting and are concerned by the systems and their outcomes (Diakopoulos, 2015). In the questionnaire, three questions support this variable, Q12_T1, Q5_T2 and Q16_T3. The first question supports the factor of understandability, the second the interpretability and the last the observability of the system.
Experience in IT sector:	As already debated in chapter 2.2, IT expertise does have an influence on the awareness and trust of the user (Micocci M et al., 2021). In the study of Gran et al. the different levels of education had a great influence on their awareness level (Gran et al., 2021). Hence, this variable tests if a specific education and experience in the IT sector does play an influence. The code of this question is Q48_it.
Dependent variables	
Self-reported algorithmic awareness:	On a 5-point Likert scale the respondents needed to indicate what level of awareness they have when it comes to personalized algorithmic platforms. For reference, the questions code is Q3_A1. Before answering this question, they got shortly introduced to the terminology of algorithmic platforms. The question was taken from the research of Gran et al. (Gran et al., 2021).
Self-efficacy:	This variable indicates the respondent's imagination on how well certain tasks are executed and performed (Hu et al., 2021). It is influenced by the FEAT-factors and is measured by three questions, with the codes Q21_S1, Q18_S2 and Q23_S3.
Trust:	With this variable, the confidence, belief and/or hope in algorithmic platforms and their decisions is expressed (Fast & Jago, 2020). It is influenced by self-efficacy and influences the awareness of mitigation measures. The variable is determined by three questions, Q17_TR1, Q22_TR2 and Q19_TR3.
Awareness of mitigation measures:	Awareness of algorithms and their unethical behavior helps to limit their wrongdoings(Gran et al., 2021). Therefore, knowing about potential mitigation measures against unethical algorithmic platforms can improve the AI environment and create a greater public debate (Costanza-Chock et al., 2022). In this study, three questions were elaborated on the awareness of mitigation measures. The first one, Q20_M1, asked what they think about audits being useful for discovering unethical behavior. The second, Q24_M2, tested the knowledge of the respondents, because they needed to write down any mitigation measure against unethical artificial intelligence that came to their mind; this question was optional. In the third question, Q25_M3, the respondents needed to select from 18 responses the ones they thought were potential mitigation measures. The ideas for the mitigation measures were taken from the literature (chapter 3). However, of the 18 responses, two were invented to test if the respondents understood what mitigation measures were and/or if they are attentively reading the questions.