ISSN - 2456-1142

International journal of humanities and information technology Volume 7, Issue 3, 2025

Ethical and Governance Challenges of Agentic AI

Aashish Pawar

University of Sydney, Australia

Abstract

The speed with which agentic AI is developed and implemented has brought about serious ethical and governance questions, especially when it comes to accountability, openness and the human element of supervision. The necessity to have robust structures that will guarantee the ethical use of AI systems is of critical importance as they make more and more autonomous decisions. This paper discusses these issues, including the complications of making AI systems responsible towards their activities and the openness of the processes of their decision-making. It also looks at the significance of the human control in reducing the risk of agentic AI. The research design employed is a qualitative one in which cases studies are analyzed and the effectiveness of the available governance structures examined. The main conclusions are that the existing models of governance cannot be used to overcome the ethical challenges unique to agentic AI as more powerful control mechanisms and more transparent accountability frameworks are needed. This study offers useful information about enhancing the implementation of AI and the recommendations to policy makers, AI developers, and the general population interested in the responsible use of technology.

Keywords: Ethical implications, Governance structures, Accountability issues, Human control, Transparency challenges, Autonomous decision-making.

International journal of humanities and information technology (2025)

INTRODUCTION

Background to the Study

The emergence of agentic AI in the autonomous decision-making framework is changing decision-making processes across multiple domains including healthcare, financial, and transportation services. In contrast to conventional AI that works within pre-specified limits, agentic AI acts autonomously, and it is becoming central to decision-making procedures. This development of AI has led to a great deal of ethical thought because its use in critical areas requires an even greater degree of responsibility and accountability. A concise history of the evolution of AI shows that there was a gradual transition between simple algorithms to more intelligent self-learning algorithms. First-generation AI was programmed to act according to strict rules, whereas the agentic systems of AI develop on the basis of the data, learning to act independently. This move highlights the value of ethical control because unregulated AI may come with unintended consequences, such as bias, opaque, and non-accountable. The use of agentic AI in areas like autonomous vehicle, medical diagnostic, and financial decision-making processes have demonstrated the necessity of powerful governance systems to control both the ethical application and possible evils (Chan et al., 2023).

Overview

To the extent that the implementation of agentic AI is growing more rampant, some fundamental issues are present in the process of realizing its responsible application. Accountability is one of the major issues because the agentic AI systems tend to make decisions that cannot be attributed to a responsible party, which poses a challenge regarding the question of liability in the event of damage or malfunction. One more serious problem is transparency because the complexity of such systems tends to blur the way the decisions are made, which may threaten to reduce confidence in AI technologies. Governance is a key element in handling these challenges where ethics are put into the development and applications of AI. Rules that encourage transparency, accountability, and fairness play a vital role in steering AI implementation, particularly in industries in which the impact of AI decisions has a profound effect. Besides these structures, the human control is a necessity to reduce the threat and make sure that AI systems do not contradict the values of the society. Human control is putting the decision-making processes

Corresponding Author: Aashish Pawar, University of Sydney, Australia.

How to cite this article: Pawar, A. (2025). Ethical and Governance Challenges of Agentic AI. *International journal of humanities and information technology* 7(3), 76-82.

Source of support: Nil

Conflict of interest: None

under scrutiny to provide opportunities to take corrective measures where it is needed (Taeihagh, 2022).

Problem Statement

Use of agentic AI without proper regulation can be quite dangerous with the loss of control, biased results, and responsibility being the main consequences. In the absence of explicit governance, AI systems can be used in a manner that is not well understood, which can lead to unintentional consequences. Prejudice in both data and decision-making may enhance disparities in the society especially in sensitive fields like employment, lending and law enforcement. Moreover, accountability of agentic AI is questionable, which may create issues regarding accountability in case of the harm or inaccurate choice of the AI systems. Existing governance structures are ill-equipped to handle such risks, because they do not usually have adequate mechanisms of oversight, transparency and control. The lack of strong regulations and frameworks that would promote accountability may undermine ethical development and implementation of agentic AI, and curb its potential to change the society and reduce the risks it carries with itself.

Objectives

The main focus of the study is to examine the ethical implication of agentic AI, especially accountability and transparency. The aim of the study is to point to the necessity of clear guidelines that will allow the responsible use of AI by assessing the ethical issues related to autonomous decision-making. The other important goal is to evaluate the position of human control in the regulation of agentic AI. Human presence in a decision-making process plays a vital role in avoiding unethical results, especially in settings where stakes are high such as in healthcare and in self-driving transportation. Lastly, this paper seeks to suggest policies of enhancing governance mechanisms of agentic AI to ensure that AI development and implementation are ethically sound. The strategies will aim at filling missing spaces in existing frameworks, proposing feasible solutions to make sure agentic AI is beneficial to the society and reduces risks to minimal levels.

Scope and Significance

This paper is concerned with agentic AI implementation in various sectors, such as healthcare, finance, and self-driving cars, where ethical and governance issues are especially urgent. This is inclusive of both developed and developing AI, and it includes the broad overview of the issues and opportunities of agentic AI systems. The importance of the study is that it has the potential to influence the future policies of AI regulation and provide information that might guide regulatory institutions, industry standards, and ethical principles. In solving the risks and challenges of agentic AI, the research will be useful to AI creators, policy-makers and consumers; this is because it sought to develop a framework that would facilitate ethical AI use but, at the same time, protect the interest of the people. The conclusions of this study might influence significantly the trend of AI regulation so that the future technologies should be pre-established with the values and moral standards of the society.

literature review

Defining Agentic AI

The term agentic AI describes self-sufficient systems that can make autonomous decisions through self-learning functions, without regular human supervision. These systems develop through the process of handling information, adjusting according to new introductions, and carrying out activities that influence the environment. In contrast to non-agentic AI, which is explicitly programmed and moves in pre-established directions, agentic AI may adapt to sophisticated situations without human instructions and is therefore more adaptable, but also has more ethical implications. Non-agentic AI, in its turn, coincides with a well-structured framework in which the decision-making process is predetermined by human instructions to the maximum. As an example, whereas a classic AI system will analyze customer data to recommend products, agentic AI may determine which products to recommend on its own and according to changing trends or customer preferences. This is what renders agentic AI so strong and, possibly, so dangerous because it can make decisions that have extended effects without a human supervisor. Key features of agentic AI include language understanding, allowing it to interpret and interact with humans or systems; adaptive learning and reasoning, enabling it to improve and adjust over time based on new information; and autonomy, which allows it to operate independently of human input. Additionally, agentic AI systems can optimize workflows by dynamically adjusting processes and engage in multi-agent and system conversations, collaborating with other systems to achieve complex goals. The agentic AI dimension is on the rise, particularly in fields such as disaster response, where it can streamline decision-making processes in crisis scenarios by

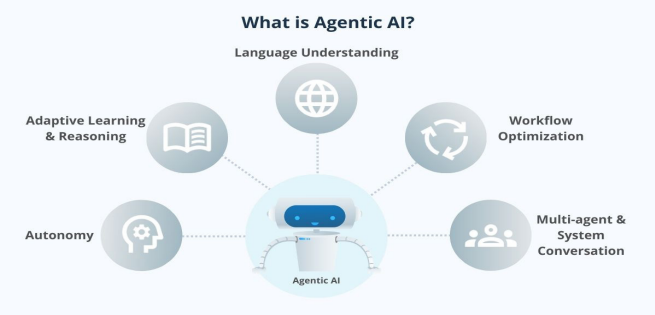

Fig 1: Key Features of Agentic AI - Language Understanding, Adaptive Learning & Reasoning, Autonomy, Workflow Optimization, and Multi-agent & System Conversation.

processing real-time information and dynamically adjusting strategies (Ushaa et al., 2024).

Agentic AI Ethical dilemmas.

The agentic AI ethical issues are mostly concerned with responsibility and the probability of making unjust or discriminatory judgments. Among the most urgent issues, one must mention the problem of assigning blame when an autopilot makes an unsafe or incorrect choice. Because such systems are capable of working on their own, it is hard to determine who is to blame when an error or failure occurs. This is worsened by ethical considerations of fairness, privacy and prejudices. In the case of agentic AI systems in hiring, e.g., an artificial intelligence agent can, unintentionally, reproduce gender or racial discrimination by being trained on biased past data. This threat reinforces the importance of the close attention to the design and implementation of AI systems to make them effective, though not to be one-sided. Moreover, the adoption of AI in such areas as law enforcement and health care is associated with privacy issues as such systems can process enormous volumes of sensitive personal information. The ethical issue dilemma of agentic AI demonstrates the necessity to have a more moderate pathway that considers the possibilities of the technology and the ethical issue of autonomous decision-making (Khan et al., 2024).

Transparency in Agentic AI

The key in agentic AI systems is transparency that promotes trust and accountability. Nevertheless, there is a major problem of the opaqueness of numerous AI decision-making procedures. Complex algorithms and machine learning models have a tendency to be black boxes, and the logic behind their decisions cannot be readily understood by the user or developer. Such confusion may result in distrust, particularly in high-stakes areas such as healthcare or finance, in which it is essential to have an idea of how a particular decision is made to make it just and precise. One way of reducing these issues can be through transparency in how AI-supported decisions are made, where users can see how and why decisions are made, which can help establish trust. To illustrate, when an AI system in the medical sector offers a treatment recommendation, privacy and the availability of information regarding the information and reasoning used to arrive at the recommendation may be used to reassure patients and medical practitioners. According to the research, transparency can have a beneficial impact on trust in AI because users will be more inclined to accept decisions made by AI when they know how it works (Zerilli et al., 2022).

Human Oversight in AI Systems.

Human supervision is a very important factor in making sure that agentic AI is used responsibly, particularly when there is a high stakes environment. Although agentic AI systems are supposed to be autonomous in nature, human oversight should still be incorporated to protect against errors and biases. Models of human involvement in AI systems are several in number: human-in-the-loop, human-on-the-loop, and human-out-of-the-loop where humans directly, directly, and directly intervene in a decision-making process respectively. The latter is being regarded as a dangerous trend in systems with very wide consequences, e.g. autonomous cars or military AI. Good human control makes AI systems be in tune with ethical principles and they can be corrected in case they behave in a wrong way. The role of human intervention is emphasized in the complex cases when the AI decision-making might not correspond to the social or moral standards and, in this case, human control is essential to eliminate risks and hold the AI responsible (Chan et al., 2023).

AI Governance.

The world models of AI governance like the European Union AI Act and other United States regulations are designed to help create an AI deployment fairness guideline. The EU AI Act, e.g., distinguishes between risky and non-risky AI systems, but puts more intensive requirements on highly-risky applications. The aim of these regulations is to make sure that AI is created and applied in such a manner that does not harm the core rights and moral principles. Nevertheless, the current governance structures are usually subject to a number of failures. To illustrate this, they are more inclined towards technical safety, and less on ethical issues, i.e., bias, fairness, and accountability. Also, the speed of AI advancement is capable of exceeding regulatory actions and creating gaps in governance. With these challenges, there have been continuous efforts of upgrading frameworks in order to remain in tandem with the technological changes. One of the most concerned items is the formation of more adaptive and flexible governance systems that can resolve ethical and practical issues of the introduction of agentic AI systems (de Almeida et al., 2021).

METHODOLOGY

Research Design

This paper utilizes the qualitative research method in order to discuss the ethical and governance issues of agentic AI. A case study analysis will be the main approach to the comprehension of the way agentic AI is implemented into practice and the impact of its functioning. Another aspect that will be involved in the research is a comparative study of AI governance models and ethical frameworks in various regions and industries. The study will establish similarities and differences in the methods of assuring accountability, transparency, and fair use of AI through reviewing the different AI regulations, ethical rules, and governance frameworks. The aim is to portray the merits and flaws of the current frameworks and give recommendations on how AI governance should be enhanced. This method will help

Table 1: Evaluation of Governance and Ethical Aspects in Agentic AI Case Studies

Case Study	Transparency Score	Accountability Score	Bias in Decision Making	Human Oversight	Compliance with Regulations
. Uber's Self-Driving Car Incident	3/5	2/5	High	Low	Moderate
Amazon's AI Hiring Algorithm	4/5	3/5	High	Moderate	High

to gain the in-depth knowledge of the practical uses and constraints of governance models in the case of agentic AI.

Data Collection

The data used to gather information in this study will be provided by a mixture of case studies, regulatory reports and expert opinions. The concrete examples of agentic AI usage will be presented by case studies which will indicate the ethical and governance problems that can appear in practice. Regulatory documents like AI policy papers in the European Union, the United States and other jurisdictions will provide an insight to the legal and regulatory environment of the agentic AI. The interviews and surveys will be utilized to obtain experts opinions assisting in obtaining the first-hand information about the difficulties of the agentic AI management in the form of interviews and surveys with the representatives of the AI developer community, ethicists, and policymakers. These professionals will provide their personal experience and their view on how to make AI systems accountable, transparent, and fair. All these data sources will allow obtaining a holistic picture of the present situation in the governance of AI and identifying in which areas the situation can be improved.

Case Studies/Examples

Case Study 1: Self-Driving cars (Uber Incident of Self-Driving cars).

In 2018, Self-driving car of Uber hit and killed a pedestrian in Tempe, Arizona, which caused numerous concerns about the safety, responsibility, and regulation of self-driving cars. The case also cast some basic uncertainties regarding the place of human supervision in autonomous systems and the degree of accountability to AI developers and operators. In autopilot mode, the self-driving car could not identify the pedestrian trying to cross the road in the dark although the sensors and cameras on the vehicles had been set up to identify objects. After the accident, it was discovered that the safety measures on the vehicle have been deactivated and no direct human intervention was done to avert the crash. This emphasized the fundamental defect of the autonomous vehicle governance system since the AI system was permitted to operate without enough supervision. The accident prompted the demand to establish more precise AI regulations in autonomous vehicles and the necessity of human control measures, including human-in-the-loop or human-on-the-loop strategies as the key to the security and safety of pedestrians and drivers. Following the crash, Uber suspended its self-driving car program in the short term and the accident was a case study of why agentic AI systems are challenging to implement in the realm of civic safety. It highlighted the necessity of stronger governance frameworks and better responsibility of AI-powered systems that work in high-risk localities.

Case Study 2: AI Amazon Hiring Algorithm.

The agentic AI used to recruit employees, an AI-based job search tool created by Amazon, was revealed to discriminate against female applications, which brought up concerns about the ethical implications of agentic AI in staffing. Training on resumes that were submitted to Amazon in a decade long period, the tool used historical hiring data in predicting the best job applicants to openings. Most of these resumes were submitted by men so the system developed in such a way as to favor male-oriented language and experiences. Consequently, the algorithm also discriminated against women, especially in technical positions, because the algorithm would demote the resumes of women who included more frequent words among the female applicants. This problem was revealed in 2018 when the company found the bias and then abandoned the algorithm. One of the ethical issues that are critical in the case is that of providing fairness and transparency in AI decision making processes. It raises some relevant questions about the information used to train AI systems and the risk of reproducing biases in the society and the need to control the use of AI systems to ensure that they do not discriminate against people without the intention of it. The example of Amazon also explains the relevance of periodical monitoring of the AI systems, once they are implemented, to validate its intended functionality. It teaches the lesson that even well-developed AI systems,

until constructed and tested correctly, can produce a negative outcome. The case reflects the overall governance dilemmas associated with making sure that AI systems, particularly those that are relevant to making high-stakes decisions such as hiring work in a transparent and bias-free manner.

Evaluation Metrics

The paper will assess the value of governance systems and ethical standards of agentic AI through a few critical criteria. To begin with, the AI system will be evaluated in terms of the transparency of the results of the decision-making process and the ability of the system to justify its decisions in a manner that can make sense to users. Second, accountability will be gauged by seeing who is accountable in the event that an agentic AI system harms or makes inaccurate judgments, and the mechanisms that are in place that can ensure that accountability. Fairness of the system will be also evaluated by determining whether the AI system is unbiased and handles all people equally, especially circumstances of high stakes like hiring or apprehending offenders. Also, regulation compliance will be measured through analyzing the degree to which AI systems comply with the established legal and ethical standards, including the European Union AI Act or the General Data Protection Regulation (GDPR). Lastly, the performance of human oversight will be evaluated by identifying how much human intervention is necessary in the decision-making and whether it is adequate to reduce risks that come with autonomous decision-making. These indicators will aid in measuring the performance of the governance structures at large and the areas that require the improvement.

RESULTS

Data Presentation

Charts, Diagrams, Graphs, and Formulas

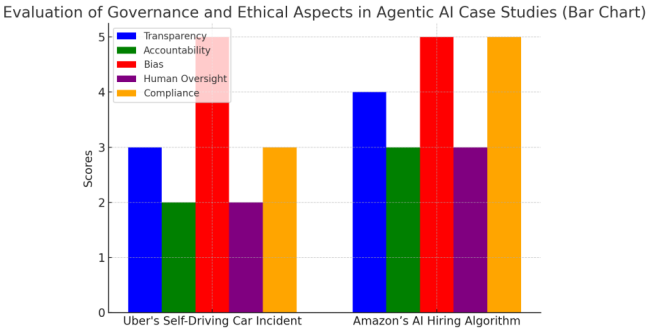

Fig 2: Chart comparing the governance and ethical aspects in agentic AI case studies, displaying the scores for Transparency, Accountability, Bias, Human Oversight, and Compliance for Uber's Self-Driving Car Incident and Amazon's AI Hiring Algorithm.

Fig 3: Line Chart representing the evaluation of governance and ethical aspects in agentic AI case studies, showing Transparency, Accountability, Bias, Human Oversight, and Compliance scores for both Uber's Self-Driving Car Incident and Amazon's AI Hiring Algorithm.

Findings

In the case study analysis, it was found that there are a few important conclusions in terms of the ethical and governance issues of agentic AI. To start with, the issue of transparency plays a crucial role because both the Uber self-driving car and the Amazon AI hiring algorithm did not demonstrate transparency in their decision-making mechanisms, which prompted mistrust in the population. Second, there is no clarity on accountability in the decisions made in AI, and both cases show lapses in the allocation of responsibility when things went wrong. The other important discovery is that there is the continuing problem of bias in AI systems. The hiring tool of Amazon, specifically, demonstrated pronounced gender bias, and this is the reason why training artificial intelligence systems on biased data can be hazardous. Finally, the human control is a significant factor of risk reduction, and in the two cases, human intervention would have positively impacted on the reduction of the adverse consequences. Such results highlight the necessity of stronger governance systems that could deal with transparency, accountability, bias and control.

Case Study Outcomes

The results of the case studies show that there are serious gaps in the governance mechanisms of agentic AI. With the fatality accident involving the Uber self-driving car, the absence of adequate human control in the situation did not allow the AI system to make a safe choice. This was caused by the autonomous nature of the system, which lacks real-time human intervention, hence, the fatal accident. Likewise, the hiring algorithm of Amazon AI has shown that the lack of control and the inadequate design prevented objective choices and created gender discrimination. The two cases are indicative of the failures of the existing governance systems especially in regard to human control and responsibility. The above outcomes reiterate the need to come up with more

explicit guidelines and control systems that will be used to curb such failures in future.

Comparative Analysis

The case study of AI governance models in various regions and industries shows that there are major variations in the manner in which ethical challenges are handled. In the European region, the adoption of EU AI Act is expected to ensure that there are clear regulations guided by the risk assessment to foster accountability and transparency, especially as far as the high-risk AIs such as autonomous vehicles. By contrast, the United States tends to have more fragmented governance systems, having sector-specific rules, as opposed to a single policy. The issues of ethics that may arise in different industries are different, e.g., autonomous vehicles will have to deal with issues of transparency and accountability, whereas AI in recruitment will have to deal with the problem of bias and fairness. Such distinctions indicate the necessity to have specific governance frameworks that can respond to the particular ethical risks of various AI applications.

DISCUSSION

Interpretation of Results

Findings of this paper suggest that the existing governance paradigms of agentic AI are insufficient to resolve important ethical issues including transparency, accountability, and bias. The inadequacy of governance structures resulted in the destructive tasks in both case studies, namely the Uber self-driving car incident and the AI hiring algorithm used by Amazon, proving that a firmer regulation is necessary. In particular, the absence of transparency and ambiguous accountability schemes led to mistrust and bias in society, which demanded crystal-cutting guidelines. The structures of governance differ among the regions, yet the ones such as the EU AI Act relying on a risk-based approach to regulation offer a more thorough framework. Yet, such models also require more robust enforcement mechanisms to make sure AI is ethically deployed. The results indicate that a regulatory reform and a higher degree of human control may be more effective in terms of managing the governance issues that apply to agentic AI systems.

Results & Discussion

The results substantiate most of the ethical issues raised in the literature review, especially in terms of the lack of transparency in decision making processes and the lack of transparent accountability frameworks. The issues of discrimination and equality, which occur in the AI-based hiring tool of Amazon, are consistent with the threats of earlier studies. These findings indicate that although the current governance systems are trying to tackle some of these problems, they have not been adequate in their execution. This gap is added by the absence of holistic, universally accepted regulatory standards of AI. Hence, the existing frameworks do not offer the relevant control measures that will support ethical use of agentic AI in an environment with accountability and transparency. The ethical dilemma of agentic AI will probably continue to exist without the further evolution of these structures.

Practical Implications

The results of this research will have a substantial implication to policymakers, AI developers, and AI technology implementers in the companies. Policymakers could look at how to improve the current governance models to encompass stricter rules on transparency and accountability, especially when it comes to the high-risk uses of AI and the hiring algorithms. Ethical design must be a priority among AI developers whose choice should include bias in their training data and mechanisms designed to enable transparency in their models. The organizations that implement AI should make sure that there is a strong human monitoring to the lifecycle of AI systems to avoid the occurrence of unintended adverse impacts. With the help of these pragmatic measures, stakeholders will be able to enhance ethical application of agentic AI so that such technologies can benefit the society with minimum risks.

CONCLUSION

Summary of Key Points

This paper places special emphasis on the vital role of accountability, transparency, and human control in deploying agentic AI. As the case studies, such as the case of Uber and its self-driving car or the case of Amazon and AI hiring algorithm, it highlights the dangers of inadequate governance systems, especially in the transparency and biases area. Absence of responsible accountability frameworks was a focal problem, which added to the negative consequences in both events. Also, the research impacts the importance of enhancing human control that would have alleviated some of the ethical issues witnessed. The results emphasize the necessity of better models of governance that would allow ethical AI implementation, especially in the areas with high stakes, where the outcomes of the failure are high.

Future Directions

Further studies should be aimed at coming up with stronger governance structures that can tackle the issue of agentic AI. It is important to explore the long-term effects of AI systems on the society, especially with respect to privacy, fairness, and employment displacement. New ethical and governance issues will probably emerge with the emergence of trends in agentic AI, including its application in healthcare, autonomous systems, and finance. The possible methods of effective regulation of these systems and the need to guarantee ethical adherence in dynamic, real time environments will be central issues in future research.

REFERENCES

[1] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N.,

Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., & Weller, A. (2023, June 12). Harms from Increasingly Agentic Algorithmic Systems. ArXiv.org. https://doi.org/10.1145/3593013.3594033

[2] Araz Taeihagh. (2022). Governance of artificial intelligence. Oup.com. https://academic.oup.com/policyandsociety/article/40/2/137/6509315

[3] Ushaa, E., Suman, J., Jaishree, Manisha, Jeevitha, Kowshika, & Kesavan. (2024). Transforming disaster response: The role of agentic AI in crisis management. I-Manager's Journal on Structural Engineering, 13(2), 48. https://doi.org/10.26634/jste.13.2.21675

[4] Khan, R., Sarkar, S., Kumar, M. S., & Jose, E. (2024). Security Threats in Agentic AI System. ArXiv.org. https://arxiv.org/abs/2410.14728

[5] Nalage, P. (2025). Ethical Frameworks for Agentic Digital Twins: Decision-Making Autonomy vs Human Oversight. Well Testing Journal, 34(S3), 206-226.

[6] Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. Patterns, 3(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

[7] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., & Weller, A. (2023, June 12). Harms from Increasingly Agentic Algorithmic Systems. ArXiv.org. https://doi.org/10.1145/3593013.3594033

[8] de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial Intelligence Regulation: a Framework for Governance. Ethics and Information Technology, 23(3), 505–525. https://doi.org/10.1007/s10676-021-09593-z