Evaluating AI Governance:

Insights from Public Disclosures

Authors

Ravit Dotan, PhD, TechBetter

Gil Rosenthal, Choir Technologies Inc.

Tess Buckley, HumansforAI, Z-Inspection

Josh Scarpino, D.Sc., TrustEngine, Assessed.Intelligence

Luke Patterson, University College London

Thorin Bristow, One World Trust

Extended Abstract

Evaluating AI Governance: Insights from Public Disclosures

This paper studies trends in AI governance based on an analysis of the public disclosures of 254 companies, such as Corporate Social Responsibility (CSR) reports and Environmental, Social, and Governance (ESG) reports. To our knowledge, this paper is the first systematic analysis utilizing public disclosures to learn about AI governance trends. Other quantitative assessments of AI governance rely on self-reporting surveys, which may be less reliable.

The analysis reveals concerning trends. First, the volume of AI ethics activities per company is typically low and the vast majority of companies did not improve over time. In fact, more companies declined than improved. Second, the analysis indicates a lack of meaningful correlation between signals of good AI governance, such as adopting AI ethic principles, and implementation activities, such as mitigating unintended bias. Third, in particular, our analysis shows that adopting AI ethics principles and similar commitments is not correlate with implementation activities. For example, 74.4% of companies with AI ethics principles and similar activities exhibited low levels of implementation.

Our findings are key to current ongoing debates surrounding AI regulation, where multiple countries promote reliance on voluntary commitments. For example, the UK's approach to AI regulation encourages reliance on voluntary commitments (Government of UK, 2023). Further, France, Germany, and Italy have expressed that they think foundation models should be self-regulated through codes of conduct and only limited sanctions (Bertuzzi, 2023). In the US, the White House announced that it has secured voluntary AI ethics commitments from several big tech companies (The White House, Sep. 2023). Similarly, Canada has recently launched a voluntary code of conduct on the Responsible Development and Management of Advanced Generative AI Systems (Government of Canada, 2023). Our findings suggest that such voluntary commitments by organizations are likely to be ineffective on their own, as we found that most companies with AI ethics commitments exhibited a low volume of corresponding implementation activity.

Our findings are also a key resource for any stakeholder who needs to evaluate the responsibility of AI governance of companies, including consumers, investors, procurement, and insurance companies. Such actors often need to evaluate companies with little or no information about their inner workings and limited technical proficiency. Therefore, they often rely on outward-facing signals, such as the existence of AI ethics principles and personnel. Whilst it is a natural tendency to rely on the governance signals companies project, our findings indicate that these signals are often misleading and therefore contribute to the phenomenon of ethics washing that clouds the AI industry.

The structure of the paper is as follows. Section 2 details our methodology. Section 3 presents the results of the analysis: 3.1 describes the prevalence of AI ethics activities, 3.2 describes the lack of meaningful correlation between the outward-facing signals and implementation activities,

and 3.3. describes the lack of progress in responsible AI governance over time. Section 4 is a discussion of these findings. It includes a discussion of the risk of mass harm, the risk of ethics washing, the prevalence and present hollowness of AI ethics commitments, a concern that risk mitigation activities are insufficiently informed, concerns that AI ethics personnel are underqualified, and topics for further research.

Abstract

This paper studies trends in AI governance based on an analysis of the public disclosures of 254 companies. The analysis reveals that many organizations exhibit signals of responsible AI governance, such as committing to AI ethics principles, but the volume of these signals per company is low. In addition, the volume of implementation activities, such as mitigating unintended bias, is especially low. Moreover, the analysis indicates a lack of meaningful correlation between signals of good AI governance, such as committing to AI ethics principles, with implementation activities. Our findings are a red flag to those who evaluate AI responsibility based on the presence of AI ethics commitments and similar governance signals. Our findings are also a red flag to those who advocate for replacing AI regulation with voluntary commitments, as these approaches may prove ineffective and augment issues of ethics washing.

Introduction

This paper studies trends in AI governance based on an analysis of the public disclosures of 254 companies, such as Corporate Social Responsibility (CSR) reports and Environmental, Social, and Governance (ESG) reports. The analysis reveals that many organizations exhibit signals of responsible AI governance, such as committing to AI ethics principles, but the volume of these signals per company is low. The volume of implementation activities, such as mitigating unintended bias, is even lower. Moreover, the analysis indicates a lack of meaningful correlation between signals of good AI governance and implementation activities.

To our knowledge, this paper is the first systematic analysis utilizing public disclosures to learn about AI governance trends. Other quantitative assessments of AI governance rely on surveys. For example, IBM (2022) and McKiensey (2022) both include statistics on the prevalence of various AI governance activities, but they rely on self-reporting surveys. Disclosures in documents like ESG and CSR reports involve personal accountability, as senior executives are personally responsible for the truthfulness of the content, adding to their reliability.

Our findings are key to current ongoing debates surrounding AI regulation. Some organizations advocate for replacing AI regulation with voluntary commitments. For example, the UK's approach to AI regulation encourages reliance on voluntary commitments (Government of UK,

2023). Further, France, Germany, and Italy have expressed that they think foundation models should be self-regulated through codes of conduct and only limited sanctions (Bertuzzi, 2023). In the US, the White House announced that it has secured voluntary AI ethics commitments from several big tech companies (The White House, Sep. 2023). Similarly, Canada has recently launched a voluntary code of conduct on the Responsible Development and Management of Advanced Generative AI Systems (Government of Canada, 2023). Our findings suggest that such voluntary commitments by organizations are likely to be ineffective on their own, as we found that most companies with AI ethics commitments exhibited a low volume of corresponding implementation activity.

Our findings are also a key resource for any stakeholder who needs to evaluate the responsibility of AI governance of companies, including consumers, investors, procurement, and insurance companies. Such actors often need to evaluate companies with little or no information about their inner workings and limited technical proficiency. Therefore, they often rely on outward-facing signals, such as the existence of AI ethics principles and personnel. Whilst it is a natural tendency to rely on the governance signals companies project, our findings indicate that these signals are often misleading and therefore contribute to the phenomenon of ethics washing that clouds the AI industry.

The structure of the paper is as follows. Section 2 details our methodology. Section 3 presents the results of the analysis: 3.1 describes the prevalence of AI ethics activities, 3.2 describes the lack of meaningful correlation between the outward-facing signals and implementation activities, and 3.3. describes the lack of progress in responsible AI governance over time. Section 4 is a discussion of these findings. It includes a discussion of the risk of mass harm, the risk of ethics washing, the prevalence and present hollowness of AI ethics commitments, a concern that risk mitigation activities are insufficiently informed, concerns that AI ethics personnel are underqualified, and topics for further research.

Methodology

1. Data Sources

The analysis is based on data collected by the company EthicsGrade. EthicsGrade collects data on the Corporate Digital Responsibility (CDR) of many of the world's largest companies, belonging to more than 100 different industries ranging from banking to automotive, to biotech.

EthicsGrade analyzed public disclosures of these companies, such as companies' ESG and annual reports. They focused on five pillars: Governance, Ethical Risk, Technical Barriers to Trust, Privacy, and Sustainability. These pillars branch into sub-topics covering finer-grained areas within AI governance, such as whether the company has AI ethics principles, whether they monitor the accuracy of their AI systems, and whether algorithmic decision-making is monitored by humans.

EthicsGrade's public dataset offers a rare opportunity to learn about how AI is governed in practice. We analyzed the most recent data available, throughout 2022. For the analysis, we selected 110 focus areas that pertain to AI governance explicitly, and we mapped them using the NIST AI Risk Management Framework, as described below.

2. Reliance on the NIST AI Risk Management Framework (AI RMF)

The National Institute for Standards and Technology (NIST) is a US agency that sits in the Department of Commerce. They are responsible for developing standards related to technology. NIST's AI Risk Management Framework (AI RMF) is one of the most well-respected frameworks for responsible AI governance.

The framework divides AI risk management activities into four pillars:

- **MAP** - Learning about AI risks and opportunities
- **MEASURE** - Measuring risks and impacts
- **MANAGE** - Implementing practices to mitigate risks and maximize benefits
- **GOVERN** - Systematizing and organizing activities across the organization

In our analysis, we sorted EthicsGrade's data into NIST's pillars, and in each pillar, we grouped activities into **"types"**. For example, one of the activity types in GOVERN is "Principles," which includes having AI ethics principles, commitments, or overarching initiatives within the company's policies. We analyzed trends in the types of activities companies reported. You can see the full list of activity types in Appendix A.

3. Division into Governance Signals and Implementation Activities

"**Governance signals**" are types of activities that external evaluators commonly use as signals of responsible AI governance. They all fall into the GOVERN pillar in the NIST AI RMF. We track the following signals:

- **Principles**: whether the company has AI ethics principles, commitments, or overarching initiatives within the company's policies.
- **Personnel**: whether the company has dedicated teams, committees, or high-level executives responsible for AI ethics oversight.
- **Thought Leadership**: involvement in industry and regulatory activism, as well as discussion of AI ethics in external communication.
- **Quality Perspective**: whether the company provides internal AI ethics training, communicates about AI ethics internally, and whether it promotes workforce diversity in AI-related teams.
- **External Assessment**: whether the company undergoes third-party AI ethics audits or assessments.

We also analyze what we call "**implementation activities.**" These are activities that implement AI ethics practices and they fall into the MAP, MEASURE, and MANAGE pillars in NIST's framework. You can see the full list of governance signals and implementation activities in Appendix A.

4. Exclusions from the analysis

Our analysis excludes information that doesn't pertain to AI explicitly:

- **Privacy**, e.g. whether the organization has a privacy policy.
- **Cybersecurity**, e.g. whether the organization has a cybersecurity strategy.
- **Displacement as a result of automation (which may or may not be AI),** e.g. whether the company communicates with the employees about automation plans and their impacts.
- **Ecology protection,** e.g. whether the company domiciles their data servers in low-carbon locations.
- **General governance,** e.g. general issue-reporting mechanisms and company-wide workforce diversification efforts.

Activities of these kinds are relevant to the responsible governance of AI, but they may be present in companies unrelated to AI. For more information about the information we excluded from the analysis, see Appendix B.

5. Limitations

Our data has four main limitations. First, it relies on public disclosures, which rely on self-reporting. Self-reports may not be fully representative of a company's state as internal actors may exaggerate positive aspects and underplay negative aspects of their company. Nevertheless, self-reported data can be acceptable in specific situations, particularly when the responsibility for accuracy and truthfulness rests directly with accountable executives. When executives are personally responsible for the information they report, they are incentivized to ensure the reliability of the data. This personal accountability acts as a safeguard, promoting due diligence and upholding the integrity of the reported information. EthicsGrade utilizes this type of data in its assessments.

Second, our data is about companies' state in 2022. Since then, the AI market has changed dramatically as a result of the generative AI boom starting from the end of 2022. However, analyzing this data still provides a rare opportunity for insight into the inner workings of AI governance in corporations and the reliability of governance signals.

Third, our data mostly represents large corporations and it centers on Western companies. Therefore, our analysis may not hold for Small and Medium Enterprises (SMEs) and non-Western companies.

Fourth, our data doesn't contain information about the companies' AI adoption. Therefore, the analysis assumes that some companies did not develop or deploy AI at all at the time. Where appropriate, we only analyzed companies that give some indication that they use AI, such as by sharing publicly their commitment to a set of AI ethics principles.

Results

1. The prevalence of AI Governance activities

1.1 Low volume of AI ethics activity, and lower volume of implementation

Of all the 254 companies in EthicsGrade's database in Q4 2022, 76% exhibited some AI ethics governance signals, and 53% exhibited some implementation activity. While these numbers may seem encouraging, the volume of activity is typically low: Of the 194 companies that exhibited governance signals, most (58%) exhibited only 1-2 types of governance signals; Of the 135 companies that exhibited implementation activities, most (70%) exhibited only 1-2 types of implementation activity.

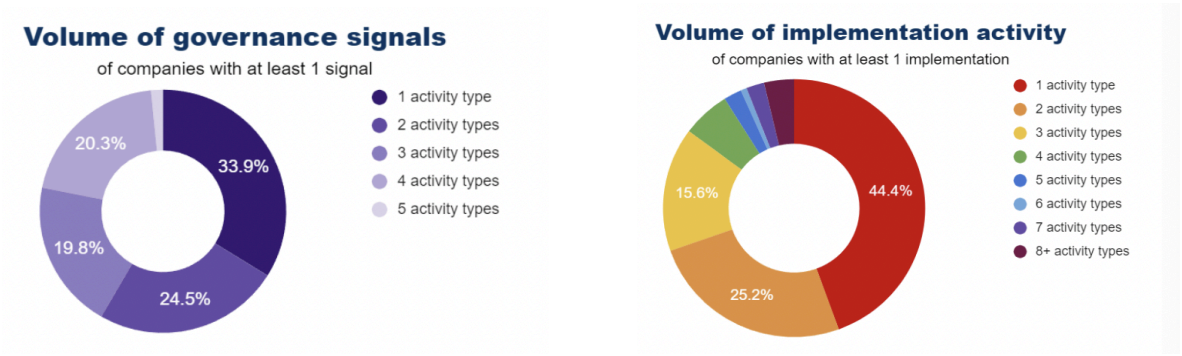

Figure 1.1: Volume of governance signals and implementation activities.

Two of the limitations of our dataset are relevant to these results. First, since our analysis is based on public data, it may deviate from what companies are doing in practice. However, the low volume of AI ethics implementation is consistent with surveys conducted around the same time as the data collected for our study (2022), such as IBM (2022) and McKinsey (2022). Both of these surveys, which are based on self-reporting surveys, reinforce the observation that companies' level of AI ethics implementation is low. For example, IBM's survey reveals that 74% of companies do nothing to reduce unintended bias.

Second, our data is mostly about large, publicly traded companies. In companies of this kind, top-down approaches are common: large companies may often start initiatives by writing a policy document. However, it is possible that smaller companies, such as startups, may be more likely to take a bottom-up approach, starting from small efforts that gradually mature into

company-wide policies. Therefore, it is possible that the ratio of implementation activities and governance signals is different in small companies.

1.2 Most common governance signals

The most common types of governance signals are Principles and Thought Leadership. 49% of all companies exhibited Principles activities, which include having AI ethics principles, commitments, or general initiatives. 47% of all companies exhibited Thought Leadership activities, which include regulatory activism, industry activism, and discussing AI ethics in external communications.

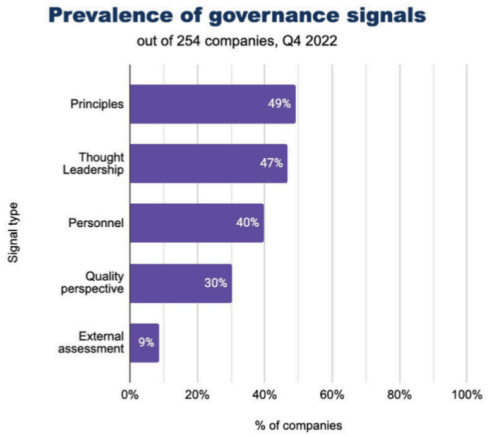

Figure 1.2: The prevalence of governance signal types in all companies.

1.3 Most common implementation activities

The most common implementation activities fall into the MANAGE pillar. 20% of all companies exhibited AI-ethics-related design and pre-review processes, which include conducting red-team exercises when developing new AI models and having operational hooks between AI ethics teams and design teams. 17% of all companies exhibited notifying users about AI, which include activities to notify users when interacting with AI and when the system has foreseeable negative consequences. We do not have information about how the notifications are provided. They may appear in terms and conditions or elsewhere.

Recall that we excluded some implementation activities from the analysis. For the full list, see Appendix B

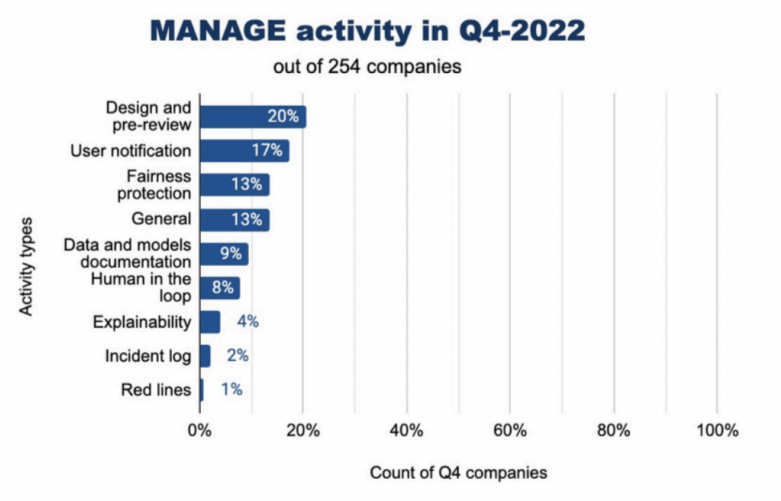

Figure 1.3: The prevalence of activity types that fall into the NIST MANAGE pillar.

2. The relationship between governance signals and implementation activities

2.1 Governance signals do not indicate meaningful implementation

Most companies that exhibit governance signals have no or low volume of implementation activities. Of all companies that exhibited at least one governance signal, 35.4% exhibited no implementation activities, and 78% presented with 2 or fewer types of implementation activities.

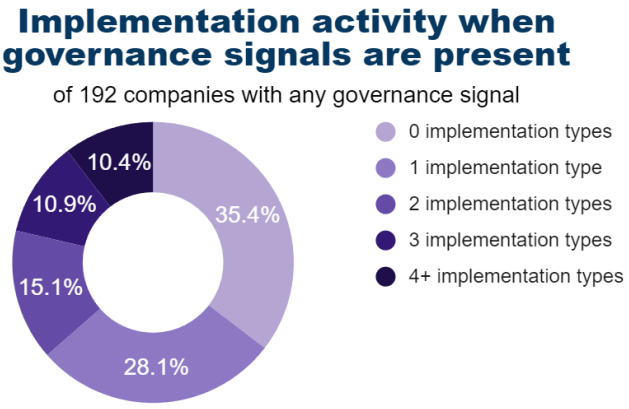

Figure 2.1a: Prevalence of implementation activity types in companies that exhibit governance signals.

In particular, most companies with Principles activities, i.e. have made AI ethics commitments, have no or low volume of implementation activities. Of the companies with AI ethics commitments, 26.4% exhibited no implementation activities and about 74.4% had 2 or fewer.

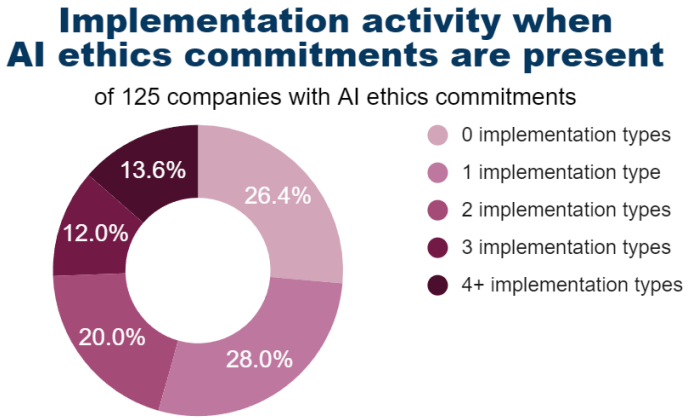

Figure 2.1b: Prevalence of implementation activities in companies that have exhibited Principles activities, such as adopting AI ethics principles.

2.2 But the more governance signals, the better

The more types of governance signals companies exhibit, the higher the average number of types of implementation activities they exhibit. For example, companies that exhibited exactly one type of governance signal exhibited an average of 0.6 types of implementation activities. Companies with five types of governance signals exhibited an average of 6.7 types of implementation activities.

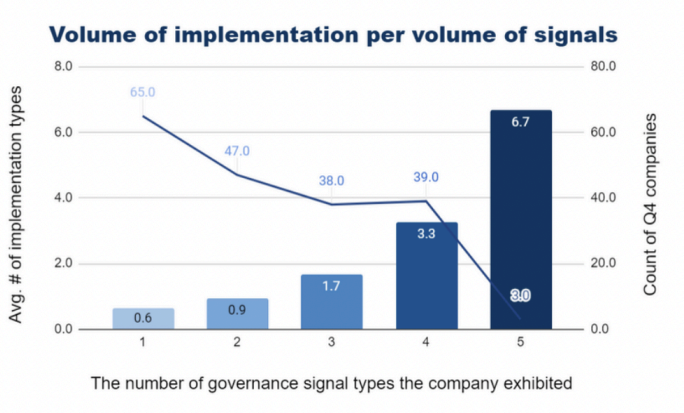

Figure 2.2a: Volume of implementation activity per volume of governance signals.

Almost half of the companies with 4 or more governance signals still exhibit low implementation: 45% of them have 0-2 implementation activity types. However, the proportion of companies exhibiting a high volume of implementation is significantly improved. The top-performing companies, with at least 4 governance signals and 4 implementation activity types, have an average of 6.8 implementation types (there are 14 companies in this category). The highest level is 11 implementation activity types, achieved by one company.

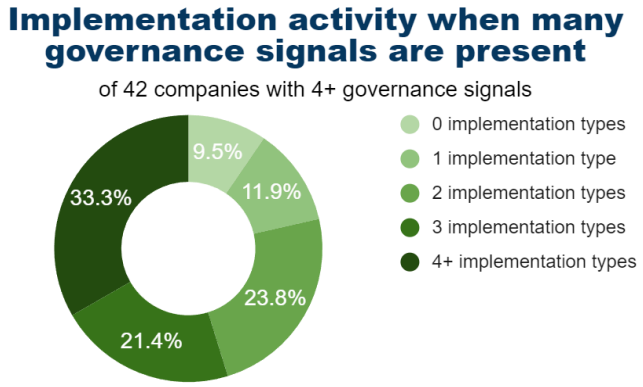

Figure 2.2b: Volume of implementation activities in companies with more than 4 types of governance signals.

2.3 An advantage of Thought Leadership

Overall, 65 companies exhibited exactly one type of governance signal in Q4 2022. The companies whose governance signal was Thought Leadership exhibited more implementation than companies whose signal was Principles or Personnel. In particular, companies whose one governance signal is AI ethics commitments performed worse: 58% of them exhibited no implementation, 37% had one, 5% had 2, and none of them had more than 2 implementation activity types.

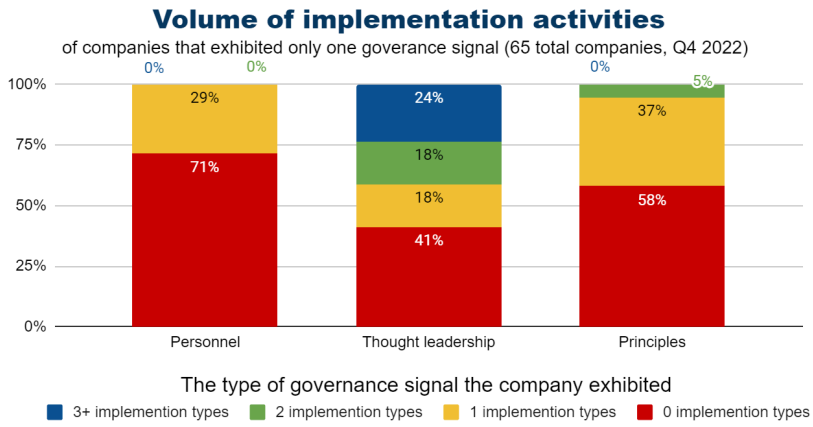

Figure 2.3: Volume of implementation activities in companies with one type of governance signal, by type of governance signal.

Having said that, note that even when the single governance activity is Thought Leadership the implementation level is low. Note that we excluded Quality Perspective and External Assessment from this analysis because the number of companies that had them as their only governance signal was too low to draw any reliable conclusions.

3. How AI ethics activities develop over time

3.1 More companies declined rather than improved, but most stayed the same

To track change over time, we compared how many types of governance signals and implementation activities each company exhibited in Q1 and Q4. In both measures, most companies (around 70%) stayed at the same level, but more companies declined than improved. In implementation activities, 17.7% declined and only 9.1% improved. In governance signals, 16.1% declined and only 13.8% improved.

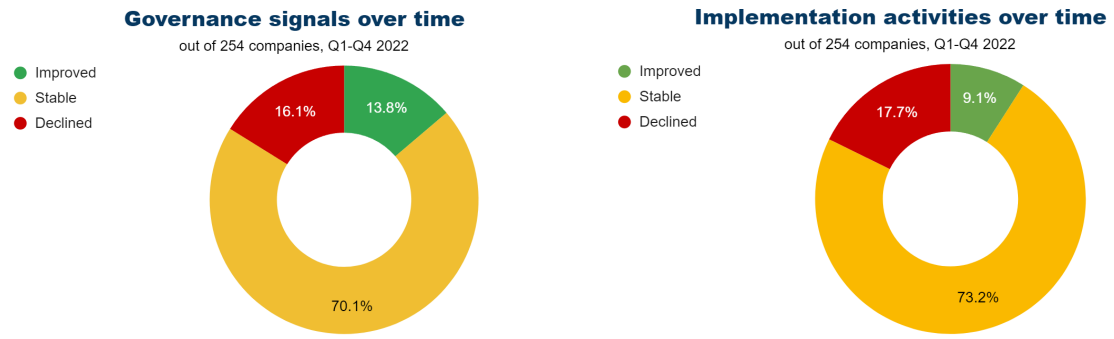

Figure 3.1a: Changes in governance and implementation activity during 2022.

Moreover, most companies fail to improve even in the presence of governance signals. For example, about 88% of companies with Principles in Q1 declined or kept the same level of implementation activities in Q4. Further, more companies with Principles declined (17.6%) than improved (11.5).

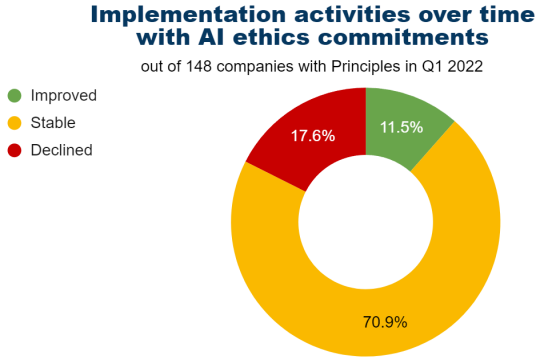

Figure 3.1b: Changes in implementation activity during 2022 in companies that had Principles activity in Q1.

3.2 Correlated with more improvement: Perspective

To learn more about what specifically characterizes the companies that improve their AI ethics performance, we looked for commonalities in governance signals between companies that had weak implementation and later improved. To that end, we isolated companies that exhibited low implementation activity (0-1 activity types) in Q1 of 2022, and then compared their implementation activities between Q1 and Q4.

We found that activities that cultivate Quality Perspective, such as AI ethics training and diversifying relevant teams, are the most correlated with implementation improvement. 18% of

companies with Quality Perspective activities improved their implementation, whereas only 10% of companies without these activities improved. The difference, 8%, is greater than the difference for other types of activities.

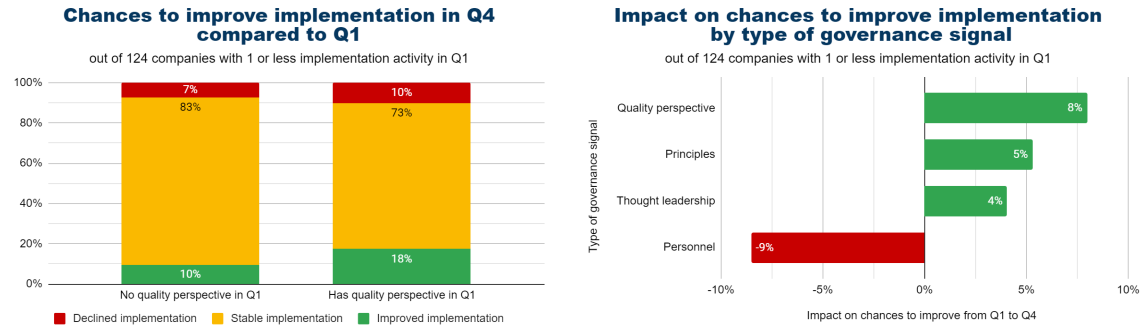

Figure 3.2: Correlation between governance signals and implementation improvement.

3.3 Correlated with less decline: Thought Leadership

To learn more about what characterizes companies that maintain strong AI ethics performance, we looked for commonalities in governance signals between companies that didn't decline in the volume of implementation activities over time. In the results described in previous sections, we noted that higher implementation volume is correlated with a higher volume of signals. Here we want to isolate governance signals that are more correlated with lack of decline.

To that end, we isolated companies that exhibited high implementation activity (2+ activity types) in Q1 of 2022, and then compared their implementation activities in Q1 and Q4. We looked for commonalities among the companies that didn't decline in their implementation in Q4. What makes a company less likely to decline?

We found that Thought Leadership activities are the most correlated with less decline. 62% of companies without Thought Leadership activities declined, whereas only 31% of companies with these activities declined. The difference, 31%, is greater than the difference for other types of activities. In particular, Thought Leadership is more correlated with lack of decline than Principles activities.

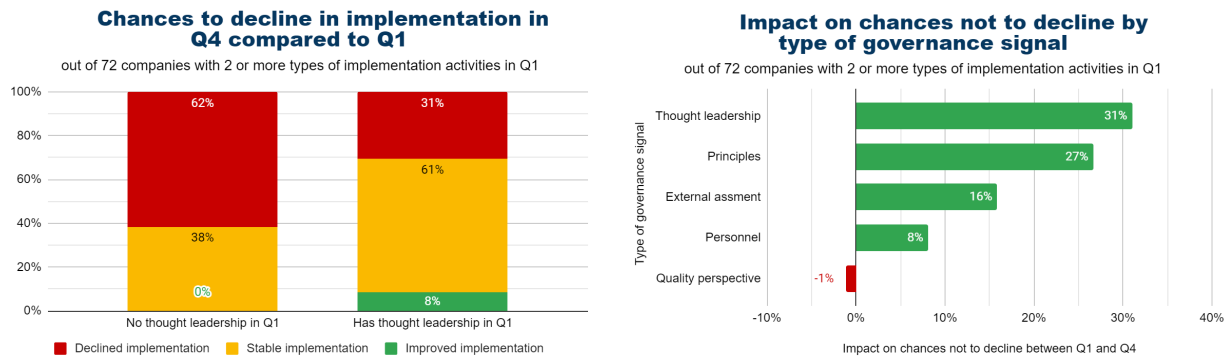

Figure 3.3: Correlation between governance signals and implementation decline.

Discussion

To our knowledge, this paper is the first systematic study of AI governance trends using companies' public disclosures. As these disclosures come with personal accountability, they may be more reliable than self-reporting through surveys. The analysis unveils concerning trends with consequences for those who regulate AI and those who evaluate AI companies.

1. Risk of mass harm: Low volume of reported AI ethics activities

We have seen that companies commonly adopt some AI ethics practices, but the volume of activities is low. In particular, the volume of implementation activities is low. Moreover, we have seen that the vast majority of companies didn’t improve their AI ethics practices during the course of 2022 and more declined than improved.

These trends are concerning. Recall that most companies in our dataset are large, publicly traded companies. The software that large companies utilize typically affects many people. If something goes wrong, it has a pronounced and widespread effect. This presents a significant risk that their AI products cause mass harm. Moreover, as AI technologies are inserted into an increasing number of a company's business processes, the potential for the scope of mass harm adjacently increases.

2. Risk of ethics washing: Lack of a meaningful correlation between governance signals and implementation

We found that governance signals generally do not correlate with meaningful implementation. For example, 78% of all companies that exhibited at least one governance signal exhibited 2 or fewer implementation activities. 35.4% of them exhibited no implementation activities. Moreover, the presence of governance signals is not correlated with improvement in implementation over time (section 2.1). Further, while implementation generally increases the more governance signals a company exhibits, it is often low even when companies exhibit many governance signals: 45% of the companies with 4 or more governance signals exhibited 2 or fewer implementation activities (section 2.2).

This result should raise the alarm for anyone who evaluates companies based on governance signals. These include consumers, investors, procurement teams, and insurance companies. Unless the company exhibits many governance signals, our data suggests that the company is probably not doing much to substantively implement AI ethics practices.

Looking ahead, our findings suggest that it is crucial to at least incentivize, and ideally require, companies to report and provide evidence on their active risk mitigation efforts in public documents used for external evaluation.

3. AI ethics commitments are popular but weak

Among the AI governance signals, commitments to AI ethics principles and similar activities are of special importance. One reason is that they are the most common kind of governance signal, with almost half the companies exhibiting some activity of this type. This finding is consistent with the explosion of AI ethics principles in all sectors. Organizations of all kinds produce such documents, including governments (e.g. The White House, Nov. 2023) inter-governmental organizations (e.g. OECD, 2023), and big tech companies (e.g. Google AI, 2023). In 2019 there were already enough of those principles for multiple review papers trying to unify them (e.g. Jobin, 2019 and Fjeld, 2019; see Dotan 2022 for a review of such papers).

Organizations may be motivated to produce AI ethics principles as a first step after which implementation would follow. In such a case, principles would represent the first indicator of substantive policies being put into practice over the coming years. However, our findings reveal a lack of evidence of a correlation between Principles activities and companies transitioning to

implementation. As discussed above, of the companies with Principles activity in Q4, 26.4% exhibited no implementation activities and 74.4% had 2 or fewer in that quarter (section 2.1). Moreover, about 88% of the companies with Principle activity in Q1 didn't improve their implementation in Q4 and many even declined. In fact, these companies were more likely to decline (17.6%) than improve (11.6%) (section 3.1).

This finding should give pause to those who push forward AI ethics principles and commitments. For example, the US White House recently signed eight big tech companies on voluntary AI ethics commitments (The White House, Sep. 2023). Similarly, Canada's Minister of Innovation, Science and Industry launched a voluntary AI code of conduct (Government of Canada, 2023). In the UK, the nation's strategic AI approach calls to rely on voluntary commitments to supplement legislation (Government of the UK, 2023). Most recently, France, Germany, and Italy have called for using voluntary codes of conduct to regulate foundation models instead of including requirements in regulation such as the EU AI Act (Bertuzzi, 2023). However, our data indicates a lack of evidence that such commitments are effective. Our analysis suggests that AI ethics principles may currently be a stronger indicator of potential ethics washing than they are of a company implementing actionable policies to mitigate the risks of inserting AI technologies into their business processes.

4. Concern: Risk mitigation activities are insufficiently informed by risk mappings and measurements

In the NIST framework that we are using, MANAGE activities pertain to implementing risk mitigation practices. MAP activities pertain to understanding the potential risks and benefits. MEASURE activities pertain to measuring risks and impacts.

Ideally, MANAGE activities should be based on MAP and MEASURE activities (i.e., risks and benefits are first mapped and measured and then mitigated based on the outcomes). For example, companies would first map how they might be impacting fairness, decide how to measure it, and use that information when planning their mitigation activities.

Our data didn't include detailed questions about MAP and MEASURE activities. For example, the only two MEASURE activity types our data tracks is whether the company monitors the accuracy of its AI models and whether they have a methodology for measuring AI risks.

Having said that, the volume of MAP and MEASURE activities in our data is extremely low (see Appendix B). This raises concerns that companies may not systematically map and measure AI risks before planning and implementing mitigation practices. For example, only 6% monitored the accuracy of their models and none reported having a methodology for measuring AI risks. These two activities are crucial for risk measurement and mitigation activities.

5. Concern: Are AI ethics personnel qualified?

Companies exhibit Principles and Thought Leadership signals the most. But who is driving these initiatives? It is notable that governance signals that provide the required expertise, employing dedicated AI ethics personnel and activities for cultivating quality perspective, are less common. The gap may be impacted by under-disclosed information about personnel and training. However, it may also indicate that AI ethics initiatives are driven by people with alternative training, such as privacy, cyber-security, and legal teams. Expertise in such areas doesn't necessarily align with the expertise required to implement a robust internal AI ethics framework. With this in mind, it could be the case that those driving AI ethics initiatives may be underqualified and therefore developing a weak base upon which to develop a robust structure of AI risk management for the future.

6. Open questions and topics for further research

6.1 Why do companies fail to move from talk to action in AI ethics?

A prominent reason may be that companies often fail to integrate AI ethics efforts into their business models. Businesses prioritize initiatives that they perceive to have a direct impact on revenue instead of implementing AI ethics, which is often thought of as a 'nice-to-have' side project or predominantly aimed at improving public relations. While the intention behind these efforts may be sincere, this approach makes it difficult for those who lead AI ethics in the company to get buy-in from senior management as well as cooperation from employees. In addition, under this approach, it is easy for AI ethics to be deprioritized whenever something arises that is perceived to align more closely with business objectives. To test this hypothesis, further research could design and test interventions to incorporate AI ethics into companies' business models.

6.2 Why may Thought Leadership be more closely correlated with implementation activities and with less decline?

Thought Leadership within responsible AI practices is a relatively strong indicator of a company moving beyond commitments and into practice and not declining over time. One potential reason is that producing Thought Leadership content creates a greater level of external expectations for the company, which increases the likelihood of implementing responsible AI practices. Another potential reason is that producing Thought Leadership requires employing personnel with a genuinely high level of expertise in AI ethics. With a workforce more attuned to and trained in issues of AI ethics, a company is better positioned to leverage internal expertise to realize substantive implementation practices.

7. Conclusion

This paper identified trends in AI governance through analysis of the public disclosures of 254 companies, based on data collected by EthicsGrade in 2022. The findings are concerning. First, while many companies exhibit some activity, the volume per company is typically low. Second, activities that often signal responsible AI governance typically do not indicate implementation activities that would impact the company's product. In particular, AI ethics principles and similar commitments typically do not correlate with implementation and could counterintuitively be a stronger indicator of ethics washing.

Appendix A - Activities classification

GOVERN	
Activity type	Content
AI ethics principles	● Existence of AI ethics principles Commitments - e.g. having AI ethics principles, committing to adopt AI industry standards ● General initiatives - e.g. general initiatives to promote public trust in AI
AI ethics personnel	● Committees and teams - e.g. AI ethics board, AI risk working group ● Executives - e.g. existence of a person responsible for tech ethics on the board
Thought leadership	● Industry activism - e.g. membership in AI ethics industry initiatives ● Regulatory activism - e.g. consult government agencies ● External communication - e.g. discuss AI ethics in external communication, provide educational materials for the public, publishing results of AI ethics audits
Quality perspective	● Internal AI ethics training AI ethics in internal communication - e.g. discussing AI ethics topics and methods ● Workforce diversity - e.g. striving for diversity in R&D teams and AI ethics committees
External assessment	● External review of principles, frameworks, and processes

Activities in MAP, MEASURE and MANAGE		
MAP Activity types	MEASURE Activity types	MANAGE Activity types
Internal input External input Input diversity Reporting mechanisms Input integration	Existence of risk measurement methodology Monitoring accuracy	Data and model documentation Design and review Explainability Fairness protection Humans in the loop Incident log Red lines User notification about AI General - implementing industry standards; putting in place measures to handle high-risk AI application

Appendix B - Excluded data

Our analysis excluded information about governance that does not pertain to AI directly:

- **Privacy**, e.g. whether the organization has a privacy policy.
- **Cybersecurity activities**, e.g. whether the organization has a cybersecurity strategy.
- **Displacement as a result of automation (which may or may not be AI)**, e.g. whether the company communicates with the employees about automation plans and their impacts.
- **Ecology protection**, e.g. whether the company domiciles their data servers in low carbon locations.
- **General governance**, e.g. general issue-reporting mechanisms and company-wide workforce diversification efforts.

While these are related to AI ethics and are important, they are too generic. When companies report that they perform these activities, there is no way of knowing whether the implementation is related to AI at all. Moreover, some of these activities are widespread and thereby unhelpful in differentiating companies' responsibility levels. For example, any company with a website is expected to have a privacy policy. Below you can find the prevalence of activities in these categories (none of them are in the MEASURE pillar):

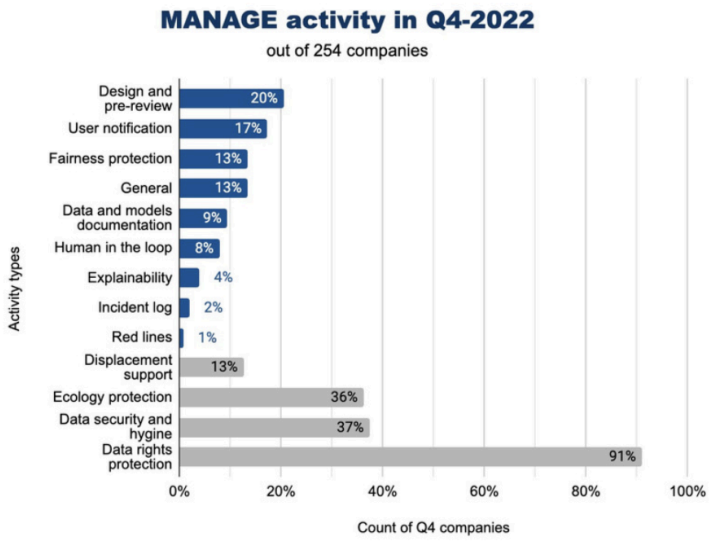

Figure B1: Prevalence of implementation activity types belonging to the NIST MANAGE pillar.

Figure B2: Prevalence of implementation activity types belonging to the NIST MEASURE pillar.

Figure B3: Prevalence of implementation activity types belonging to the NIST MAP pillar.

Bibliography

Bertuzzi, Luca. “France, Germany, Italy Push for ‘Mandatory Self-Regulation’ for Foundation Models in EU’s AI Law.” *Www.Euractiv.Com*, 19 Nov. 2023, www.euractiv.com/section/artificial-intelligence/news/france-germany-italy-push-for-mandatory-self-regulation-for-foundation-models-in-eus-ai-law/

Chui, Michael, et al. “The State of AI in 2022-and a Half Decade in Review.” *McKinsey & Company*, McKinsey & Company, 6 Dec. 2022, www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#/download//~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai%20in%202022%20and%20a%20half%20decade%20in%20review/the-state-of-ai-in-2022-and-a-half-decade-in-review.pdf?cid=soc-web.

Dotan, Ravit. “The Proliferation of AI Ethics Principles: What’s Next?” *Montreal AI Ethics Institute*, 12 Feb. 2022, montrealethics.ai/the-proliferation-of-ai-ethics-principles-whats-next/.

Fjeld, Jessica, et al. “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.” *SSRN*, 14 Feb. 2020, papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482.

Google AI. “Google Ai Principles.”, 2023, ai.google/responsibility/principles/.

Government of Canada. “Minister Champagne Launches Voluntary Code of Conduct Relating to Advanced Generative AI Systems.” *Canada.Ca*, 27 Sept. 2023, www.canada.ca/en/innovation-science-economic-development/news/2023/09/minister-champagne-launches-voluntary-code-of-conduct-relating-to-advanced-generative-ai-systems.html

Government of the United Kingdom, “A Pro-Innovation Approach to AI Regulation.” *GOV.UK*, 3 Aug. 2023, www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

IBM, “IBM Global AI Adoption Index 2022.”, May 2022, www.ibm.com/watson/resources/ai-adoption.

Jobin, Anna, et al. “The Global Landscape of AI Ethics Guidelines.” *Nature News*, Nature Publishing Group, 2 Sept. 2019, www.nature.com/articles/s42256-019-0088-2.

OECD, “OECD AI Principles Overview .” *OECD.AI Policy Observatory*, 2023, oecd.ai/en/ai-principles.

The White House, “Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by Ai.” *The White House*, The United States Government, 12 Sept. 2023, www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/#:~:text=The%20companies%20commit%20to%20publicly,effects%20on%20fairness%20and%20bias.

The White House, “Blueprint for an AI Bill of Rights.”, The United States Government, 22 Nov. 2023, www.whitehouse.gov/ostp/ai-bill-of-rights/.