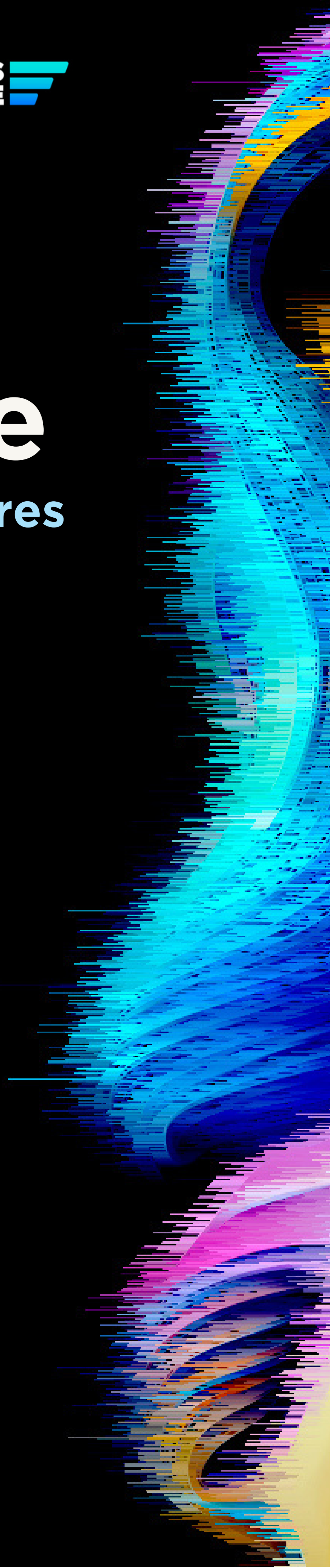
TechBetter

ETHICS GRADE

Evaluating AI Governance

Insights from Public Disclosures



Ravit Dotan
Gil Rosenthal
Tess Buckley
Josh Scarpino
Luke Patterson
Thorin Bristow

Executive **Summary**

Artificial intelligence (AI) adoption has exploded in 2023, with tools such as ChatGPT dramatically raising awareness of the potential of these technologies for commercial and personal use. In this changing landscape, it is increasingly important to evaluate organizations that develop and deploy AI systems. Do they identify their impacts? Are they managing them responsibly?

Companies often disclose information that can help answer these questions in public documents on their websites, annual reports, ESG reports, and more. For people who need to assess companies with minimal access to internal information – like consumers, investors, and procurement teams – this public information is especially valuable. They must decide whether to use, buy, invest, or otherwise support companies and products. Knowing if the company governs AI responsibly can be crucial for those decisions.

In this report, we analyze companies' AI governance based on the information they publicly provide.

We find that the volume of reported AI ethics activities is low. Moreover, we find that typical governance signals, including the existence AI ethics principles, do not correlate with implementation.

Therefore, we recommend against solely relying on signals and that companies be incentivized or required to report on their implementation activities.

About Our **Analysis**

Our analysis is based on data collected by EthicsGrade. EthicsGrade collected data about the corporate digital responsibility (CDR) of 254 companies between 2021-2022. This included data regarding AI governance, such as whether a company has established AI ethics principles and whether they monitor the accuracy of their AI systems. We analyzed EthicsGrade's data from 2022. We used the framework set by the NIST AI Risk Management Framework (NIST AI RMF), and analyzed types of activities that fall into one of the pillars of the NIST AI RMF:

- **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

Governance signals

We were especially interested in **governance signals**, types of activities that external evaluators commonly use as signals of responsible AI governance. The governance signals we tracked were:

- **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.

These signals may contribute to ethics washing if they are not accompanied by **implementation activities**, where companies take meaningful internal action to map, measure, and manage their AI risks. Our study sheds light on the relationship between governance signals and implementation activities.

Summary of **Findings**

1. Prevalence of AI ethics activities

1.1 Low volume of AI ethics activity, lower implementation

Of all 254 companies in EthicsGrade's dataset in Q4 of 2022:

- 76% exhibited AI ethics governance signals.
- 53% exhibited implementation activities.

When companies report AI ethics activities, the volume is low:

- Of the companies that exhibited governance signals, 58% had only 1-2 types of these activities.
- Of the companies that exhibited implementation activities, 70% had only 1-2 types of these activities.

1.2 Most common governance signals

- AI ethics principles, commitments, etc. is the most common governance signal (49% of all 254 companies in Q4 2022).
- Thought leadership, which includes regulatory activism, industry activism, and discussing AI ethics in external communication, is the second most common (47% of all 254 companies in Q4 2022).

1.3 Most common implementation activities

- Design and pre-review activities are the most common type of implementation activity companies exhibited. These activities include conducting red-team exercises when developing new AI models and having operational hooks between AI ethics teams and design teams.
  (20% of all 254 companies in Q4 2022).
- Notifying users when they engage with AI or when the AI system has foreseeable negative consequences is the second most common type of implementation activity.
  (17% of all 254 companies in Q4 2022).

2. The relationship between governance signals and implementation

2.1 Governance signals do not indicate implementation

- 27% of all companies exhibited governance signals but no implementation
- Of the companies that exhibited exactly one type of governance signal:
  - 58% exhibited no implementation and 28% exhibited only one type.
- Of the companies that exhibited exactly two types of governance signals:
  - 40% exhibited no implementation and 36% exhibited only one type.

2.3 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.

2.4 Thought Leadership is the governance signal most indicative of implementation activities

- 65 companies exhibited exactly one type of governance signal in Q4 2022.
- When the one governance signal was thought leadership, companies exhibited more implementation activity than companies relying on any other individual signal.

3. How AI ethics activities develop over time

3.1 More companies declined than improved AI ethics activities during 2022, but most stayed the same

Comparing between Q1 and Q4:

Implementation:

- 73.2% had the same number of implementation activity types.
- 17.7% declined in the number of implementation types they exhibited, while only 9.1% improved.

Governance signals:

- 70.1% had the same number of governance signal types.
- 16.1% declined in the number of governance signals they exhibited, while only 13.8% improved.

3.2 Correlated with more improvement: Perspective

- 18% of companies with Quality Perspective activities in Q1 improved in implementation activities in Q4.
- Only 10% of companies without Quality Perspective activities improved.
- The difference, 8%, is greater than the difference for other signal types.

3.3 Correlated with less decline: Thought Leadership

- 31% of companies with Thought Leadership activities in Q1 declined in implementation activities in Q4.
- 62% of companies without Thought Leadership activities declined within the same period.
- The difference, 31%, is greater than the difference for other signal types.

Key **Reflections**

Low volume of reported AI ethics activities

It is concerning to see the low volume of AI ethics implementation as well as the lack of any significant improvements over the course of 2022. It is also concerning to see the lack of correlation between governance signals and implementation activities.

No evidence that AI ethics principles and commitments lead to implementation.

In particular, it is notable that the existence of AI ethics principles and commitments, the most common governance signal, is not positively correlated with exhibiting implementation activities. Various organizations advocate the adoption of voluntary AI ethics commitments. These include the US and Canada, which recently launched initiatives to encourage companies to commit to AI ethics codes of conduct. It also includes the UK, whose national approach to AI relies on voluntary codes of conduct. However, our report indicates a lack of evidence that such commitments are effective.

The discrepancy between governance signals and implementation activities may contribute to ethics washing.

Given the lack of evidence for a correlation between governance signals and implementation, governance signals may mislead the public and other external evaluators. Their consumption and other choices could be impacted by neat AI ethics activities that look good but are not backed up by practices that impact the product. Many AI ethicists express concerns that ethics washing is rampant in the field of AI. Our findings are consistent with this sentiment and indicates that relying on governance signals when evaluating companies is ill-advised.

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.

To Learn More

- Come to our webinar, Nov. 1, 12-1pm ET
  Registration at www.ravitdotan.com

- Full paper coming soon!