International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-14 Issue-11, October 2025

Responsible Disclosure in the Age of Generative AI: A Normative Model for Dual-Use Risk

Fahd Malik, Muhammad Raza ul Haq

Abstract: ***The rapid growth of generative artificial intelligence (AI) systems such as large language models (LLMs) has created a profound disclosure dilemma: when should potentially dangerous models or findings be shared openly, withheld, or released in a controlled manner? Traditional norms of open science and open-source software emphasize transparency, reproducibility, and collective progress, yet the dual-use nature of frontier LLMs raises unprecedented challenges. Unrestricted disclosure can enable malicious use cases such as cyberattacks, automated disinformation campaigns, large-scale fraud, or even synthetic biology misuse. In contrast, excessive secrecy risks undermining trust, slowing scientific progress, and concentrating power in a small number of actors. This paper develops a normative model for responsible disclosure that integrates utilitarian, deontological, and virtue-ethical reasoning to justify a proportional approach rather than binary openness or secrecy. We introduce a Disclosure Decision Matrix that evaluates four key dimensions: risk severity, exploitability, mitigation availability, and public benefit of transparency. It then recommends one of three courses of action: full release, staged or controlled release, or temporary restriction until safeguards mature. The contribution is twofold. First, it provides a principled ethical framework that links philosophical justification directly to operational disclosure practices, bridging the gap between theory and governance. Second, it translates this framework into actionable criteria that policymakers, research institutions, and developers can consistently apply across evolving AI systems. By combining ethical reasoning with practical decision tools, the findings underscore that responsible disclosure in AI is neither absolute secrecy nor unqualified openness but a dynamic, proportional strategy responsive to both technological advances and societal risks.***

Keywords: Generative AI; Large Language Models; Responsible Disclosure; Dual-Use Risk; Disclosure Decision Matrix

Nomenclature:
LLMs: Large Language Models.
CVD: Coordinated Vulnerability Disclosure.
CERTs: Computer Emergency Response Teams
IBCs: Institutional Biosafety Committees.
AI: Artificial Intelligence

I. INTRODUCTION

Generative artificial intelligence (AI) has moved from research labs to being used for everyday applications at Unprecedented speed. Large language models (LLMs) such as GPT-4, Claude, and Llama 2 now power writing assistants, coding co-pilots, search engines, and customer support systems. These models are trained on massive datasets and demonstrate generalization capabilities that frequently exceed researchers' expectations. They can generate coherent text, produce functional code, and summarize technical material, making them transformative tools across education, law, healthcare, and software engineering. However, as these models become more powerful and widely available, they raise an urgent question that how should society handle the risk that these same models could be weaponized?

The challenge is rooted in dual-use risk**,** the idea that the same technology can serve beneficial and harmful purposes. Researchers have shown that LLMs can produce convincing phishing emails, social engineering scripts, and even malware that exploits common vulnerabilities [1]. In controlled settings, models have been prompted to outline steps for synthesizing hazardous biological compounds [2]. While many of these outputs still require domain expertise to execute, the trajectory of improvement suggests that barriers to misuse will continue to fall. The democratization of harmful knowledge is what makes generative AI distinct from traditional expert-only risk domains.

This creates a fundamental tension between two values, i.e., openness and security. Openness has long been a cornerstone of scientific progress. The open science movement emphasizes reproducibility, transparency, and knowledge sharing to accelerate discovery [3]. Similarly, open-source software has shown that security can improve when many researchers can inspect and patch code. Yet generative AI challenges this intuition. Once model weights are published, they can be copied and redistributed globally with no way to recall them. If those models enable malicious capabilities, such as automated vulnerability discovery or scalable disinformation campaigns, the risk becomes effectively irreversible.

We have precedents in other high-risk domains. Cybersecurity researchers developed coordinated vulnerability disclosure (CVD) as a standard for handling software flaws. Under CVD, vulnerabilities are reported privately to vendors, who are given a fixed remediation window before public disclosure. This balances the public's right to know with harm reduction.

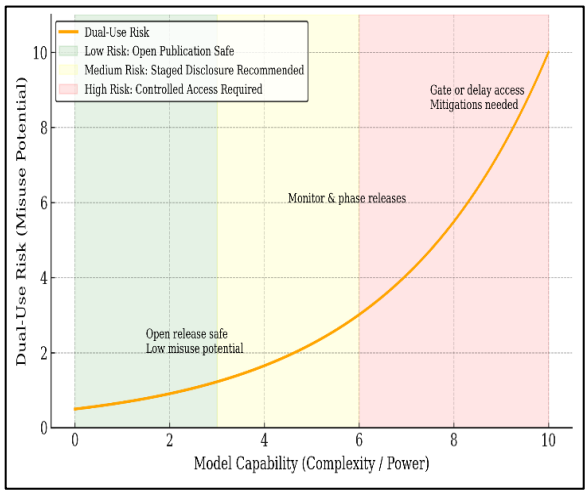
Generative AI currently lacks an equivalent, standardized process. Model release decisions are made mainly internally by private labs or research groups. Some organizations, such as OpenAI, have experimented with staged release strategies, first demonstrated with GPT-2 in 2019. The full model weights were initially withheld, with partial access granted as monitoring showed that the risk of misuse was

Manuscript received on 03 October 2025 | Revised Manuscript received on 11 October 2025 | Manuscript Accepted on 15 October 2025 | Manuscript published on 30 October 2025
*Correspondence Author(s)
Fahd Malik*, Department of Digital Business, Transformation & Innovation, IE Business School, Madrid, Spain. Email ID: fahad.agmalik@gmail.com, ORCID ID: 0009-0003-4656-0126.
Muhammad Raza ul Haq, Department of Information Technology, Zain, Riyadh, Saudi Arabia. Email ID: Muhammad.Razaulhaq@sa.zain.com, ORCID ID: 0009-0000-8614-1178.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open-access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

manageable [4]. While this approach was criticized as paternalistic, many labs have since adopted similar caution, releasing smaller or less capable models first and reserving more powerful versions for controlled settings. Conversely, open-weight releases like Llama 2 have reignited debates about whether unrestricted access is worth the potential security risks [5].

What is missing from the current landscape is a principled framework to guide these decisions. The status quo relies.



[Fig.1: Conceptual Spectrum of Dual-Use Risk and Disclosure Strategies for AI Models]

On ad hoc judgments by individual organizations, without a shared set of criteria for determining when disclosure is appropriate, delayed, or restricted. This paper argues that disclosure decisions should be grounded in moral reasoning, not just pragmatic risk assessments. Neither radical openness nor permanent secrecy is ethically defensible. Instead, we propose a normative model in which disclosure obligations are proportional to three factors: (1) the severity of potential harm, (2) the ease with which malicious actors can exploit the capability, and (3) the availability of mitigations or defences. Fig. 1 illustrates the relationship between model capability and associated societal risk.

The rest of this paper proceeds as follows. First, we review the literature on dual-use risk, vulnerability disclosure norms, and AI governance proposals. We then develop a normative analysis using utilitarian, deontological, and virtue-ethical frameworks to clarify the moral duties of AI developers. Based on this analysis, we introduce a decision matrix that operationalizes proportional disclosure and apply it to historical and contemporary cases, including GPT-2's staged release, vulnerability-generating outputs, and open-weight model debates. Finally, we offer policy recommendations for developers, researchers, and regulators seeking to balance openness, innovation, and security in the age of generative AI.

II. LITERATURE REVIEW

The idea that technology can simultaneously empower and endanger society has long shaped the ethical discourse on innovation. The "dual-use" dilemma, where a single advance carries both beneficial and harmful potential, is not new. The nuclear era was one of the earliest instances where openness in science collided with catastrophic risk. The Manhattan Project produced a scientific triumph but also introduced existential danger, prompting decades of nonproliferation treaties and institutionalized secrecy.

In [6], the authors find that these frameworks significantly reduce the social cost of software insecurity by aligning incentives among researchers, vendors, and users. Together, these studies provide a governance precedent, showing that transparency and security need not be mutually exclusive if procedures are carefully designed.

The AI domain has rapidly entered its own dual-use debate as model capabilities have scaled. In [7], the authors produced one of the first comprehensive risk maps, outlining malicious applications such as automated disinformation, cyber offence, and mass surveillance, and recommending red-teaming and staged release strategies. More recent work has demonstrated concrete misuse scenarios. In [8], authors empirically showed that large language models (LLMs) can be prompted to generate step-by-step exploit code and even novel vulnerability patterns, suggesting that access controls and safety layers are insufficient if deployment practices remain open by default. Similarly, authors in [9] catalogued potential misuse harms of foundation models and argued that model governance must be proactive, given the difficulty of retracting open weights. These results shift the debate from theoretical possibility to demonstrable risk, making the question of disclosure governance urgent rather than speculative.

Recent developments highlight that the risks of generative AI misuse are tangible rather than speculative. Regulators have increasingly raised concerns about AI-enabled voice cloning scams and taken formal steps toward oversight [10]. Similar actions include classifying AI voice impersonation under existing telecommunications regulations, which now exposes malicious actors to legal liability [11]. In [12], authors showed through benchmarking studies that large language models remain susceptible to jailbreak and prompt injection attacks, with standardized evaluations documenting reproducible exploitation across both closed and open-weight systems. Advocates contend that releasing model weights enhances reproducibility and democratizes access. Authors in [13] argue that open releases enable independent safety research and distribute oversight capacity across the research community, countering the concentration of power in a handful of corporations. However, critics raise concerns about the risk of misuse once weights are irreversibly released. In [14], authors caution that model openness without safety infrastructure expands the attack surface for malicious fine-tuning and content generation. These tensions illustrate that openness cannot be treated as a binary but as a spectrum requiring context-sensitive decisions.

Governance proposals are beginning to catch up. The OECD AI Principles (2019) and NIST AI Risk Management Framework (2023) advocate for transparency, accountability, and proportional risk management but do not yet specify clear disclosure thresholds. In [15], the authors propose a decision-theoretic model to assess when the benefits of sharing AI research outweigh the risks, integrating offence–defence considerations into the publication decision.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

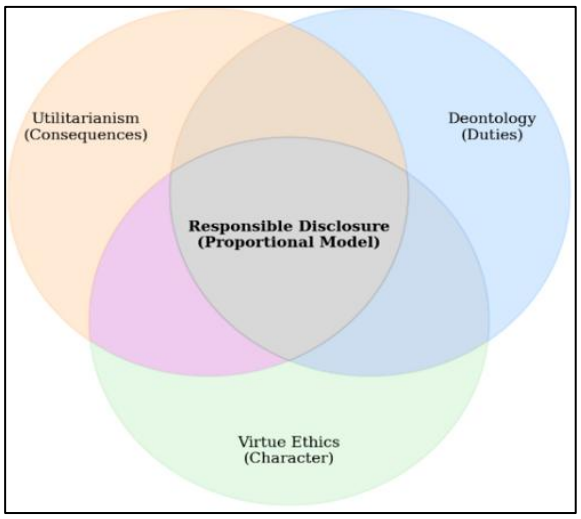
Complementing this, authors in [16] document OpenAI's staged release of GPT-2, which withheld full model weights until misuse risks could be better characterized, offering a real-world example of graduated disclosure. Such frameworks demonstrate that staged or conditional release is a viable path that preserves some openness while mitigating immediate risk. In [17], the authors explicate the precautionary principle, arguing that when potential harm is catastrophic and uncertainty is high, erring on the side of caution is ethically justified.

III. NORMATIVE ANALYSIS

A. Framing the Ethical Problem

The dual-use dilemma in generative AI is not just a technical challenge but a moral one. Deciding whether to release, delay, or restrict a model involves assigning weight to competing ethical values: the value of openness, the duty to prevent harm, and the need to respect human rights such as access to knowledge. Without a clear normative framework, disclosure decisions risk becoming arbitrary, guided only by organizational risk tolerance or public relations considerations. A principled approach is therefore essential, not only to produce justifiable outcomes but also to generate legitimacy in the eyes of stakeholders.

In this section, we examine three major moral theories as shown in Fig. 2, i.e. utilitarianism, deontological ethics, and virtue ethics, and apply them to the question of LLM disclosure. Each theory offers a distinct lens for evaluating when disclosure is obligatory, permissible, or impermissible. We argue that no single theory is sufficient on its own but that their synthesis supports a proportional disclosure model.



[Fig.2: Normative Ethical Considerations for AI Disclosure Decisions]

B. Utilitarian Lens: Minimizing Expected Harm

Utilitarianism holds that the morally right action is the one that maximizes overall well-being or minimizes expected harm. Applied to LLM disclosure, this perspective suggests that researchers should disclose in a way that produces the best aggregate outcome for all stakeholders. If releasing model weights would significantly raise the probability of catastrophic misuse — such as enabling low-skill actors to deploy ransomware at scale —then delaying or restricting release becomes ethically justified.

Quantifying harm is challenging but not impossible. One can construct a risk equation:

i. *Expected Harm = Probability of Misuse × Severity of Consequences × Population Exposed*

Under this model, high-severity but low-probability risks may still justify strong controls if the potential impact is catastrophic. Conversely, if withholding the model prevents defensive research that could reduce systemic risk, then secrecy might paradoxically *increase* expected harm. The utilitarian solution is dynamic, i.e., disclosure timing should adjust as mitigations improve, reducing the harm term and allowing eventual openness.

C. Deontological Lens: Duties and Rights

Deontological ethics emphasize that specific actions are morally required or forbidden regardless of consequences. Researchers have a duty to avoid causing harm, which implies an obligation to warn when they discover dangerous capabilities. At the same time, they have a duty to respect the rights of others, including the right to scientific freedom and the right to benefit from technological progress.

From a deontological standpoint, blanket secrecy is hard to justify. Permanent suppression of research could be seen as a violation of epistemic rights, mainly if the knowledge could be used for good purposes such as improving cybersecurity or advancing science. A deontologist might argue that researchers have a prima facie duty to disclose. Still, this duty can be overridden temporarily if disclosure would violate a stronger duty, such as the duty to protect innocent life. Importantly, deontological reasoning emphasises fairness, i.e., disclosure decisions should not favour the interests of the developer (e.g., competitive advantage) over the public good.

D. Virtue Ethics Lens: Character and Practical Wisdom

Virtue ethics shifts the focus from rules and consequences to the moral character of decision-makers. The relevant question becomes what a virtuous researcher or lab would do in this situation? A virtuous actor would display prudence by carefully weighing risks before acting, courage by disclosing when it is difficult or unpopular to do so, and responsibility by taking ownership of downstream effects.

Virtue ethics also emphasizes the cultivation of institutional culture. Organizations should reward responsible disclosure practices rather than punishing whistleblowers or incentivizing reckless releases for publicity. This lens highlights the importance of transparency in the decision-making process itself. Even if a lab withholds a model, explaining the reasoning can maintain trust and signal that the choice was made from a place of prudence rather than control.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Table I: Ethical Frameworks and their Implications for Responsible Disclosure Choices

Ethical Lens	Core Test	Disclosure Posture Under High-Risk, Low-Mitigation Conditions
Utilitarian (consequences)	Net expected harm vs. benefit	Delay or restrict until expected harm is outweighed by societal benefit
Deontological (duties)	Duties to prevent harm: duties of transparency	Prioritize duty of non-maleficence; disclose only to vetted or limited audiences.
Virtue ethics (character)	Responsible stewardship; prudence	Favour cautious, staged disclosure while signalling ethical responsibility.

E. Integrating the Lenses: Toward a Proportional Model

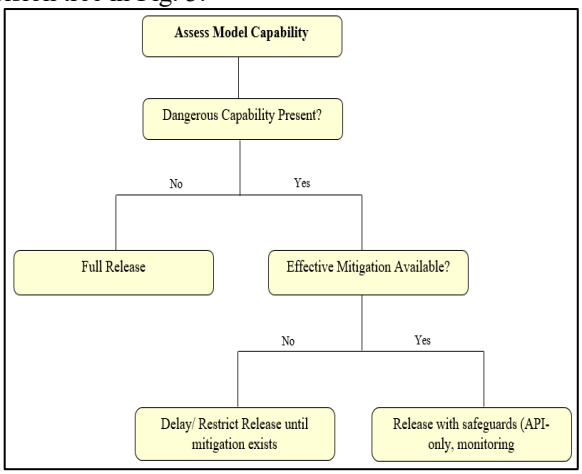
Taken together, these three approaches converge on a principle of proportional disclosure:

i. Utilitarianism tells us to minimize expected harm, justifying delays until mitigations exist.
ii. Deontology reminds us that disclosure is a moral default, to be overridden only when harm prevention takes precedence.
iii. Virtue ethics ensures that decisions are made with integrity, transparency, and long-term stewardship in mind.

The synthesis of these lenses supports a graduated response rather than a binary choice between openness and secrecy. It implies that model release decisions should be sensitive to context, scalable with risk, and revisited over time as capabilities and defences evolve. Table I shows ethical frameworks and their implications for responsible disclosure choices.

F. Operationalizing Ethical Reasoning

While moral theory provides the foundation, practitioners need tools they can actually use. We therefore propose translating these ethical insights into a decision-making workflow. The first step is capability assessment, where it is determined whether the model can produce dangerous outputs such as zero-day exploits, bioweapon instructions, or targeted disinformation. The second step is exploitability analysis, which measures how easily a malicious actor could use the model to cause harm, considering skill requirements, cost, and access barriers. The third step is mitigation availability. Check whether effective defences, such as filters, detection tools, or legal controls, exist or can be deployed in parallel with release. This workflow can be visualized as a decision tree in Fig. 3.



[Fig.3: Decision Tree for Disclosure Governance of Potentially Dangerous AI Capabilities]

A comparative overview of how disclosure strategies vary with risk level and mitigation maturity is presented in Table II.

Table II: Comparative Summary of Disclosure Strategies Across AI Risk Levels

Parameter	Low Risk / Mature Mitigation	Moderate Risk/Emerging Mitigation	High Risk / Immature Mitigation
Capability	Harmless tools, benchmarking	Code-assist, agent scaffolding	Exploit generation, disinformation vectors
Exploitability	Low	Medium	High
Mitigation maturity	Strong, proven	Partial, evolving	Weak/Untested
Disclosure	Full release	Staged/API-only	Delay/ Restrict

G. Role of Governance and Coordination

Normative reasoning also implies obligations at the ecosystem level. No single lab can fully control the risk landscape; therefore, disclosure decisions should involve coordination among multiple stakeholders. Creating neutral bodies similar to Computer Emergency Response Teams (CERTs) in cybersecurity could help mediate between developers, researchers, and governments. These bodies could evaluate risk claims, set industry-wide timelines for staged release, and prevent unilateral decisions that may harm the public good.

H. Limits of Normative Analysis

It is essential to acknowledge that moral reasoning cannot eliminate uncertainty. Predictions about misuse often rely on speculation, and overestimating risk may unjustifiably delay beneficial innovation. The model we propose is therefore not a final answer but a guide, a tool for structuring deliberation so that decisions are principled rather than ad hoc. By embedding moral reasoning into release governance, we aim to make the decision-making process more transparent, defensible, and accountable.

IV. DISCLOSURE DECISION MATRIX AND APPLICATION

The normative analysis in the previous section establishes that responsible disclosure cannot be binary. A purely open or purely secret approach fails to capture the proportionality required when balancing harm prevention, scientific freedom, and public accountability. To operationalize this insight, we propose a **Disclosure Decision Matrix**. This structured tool guides researchers and developers in deciding whether to fully release, partially disclose, or temporarily withhold model-related information. This matrix is shown in Table III.

A. Structure of the Disclosure Decision Matrix

The matrix is built on four key criteria: risk severity, exploitability, mitigation availability, and public benefit of openness. Each is rated qualitatively on a three-point scale (low, medium, high). The resulting combination

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

determines the recommended disclosure action.

B. Applying the Matrix: Case Studies

To demonstrate the utility of the decision matrix, we apply it to three illustrative cases, i.e., GPT-2's staged release, vulnerability-generating outputs, and open-weight model releases such as Llama 2.

Case 1: GPT-2 Staged Release

Table III: Disclosure Decision Matrix Mapping Model Capability Risk (Low to High) Against Mitigation Maturity (Immature to Mature)

Criterion	Low	Medium	High
Risk Severity	Misuse would cause minor harm or inconvenience.	Misuse could cause targeted harm (e.g., financial fraud).	Misuse could cause widespread or catastrophic harm (e.g., a cyberattack at scale or bioterrorism guidance).
Exploitability	Requires expert skill or significant resources.	Requires moderate expertise or access.	Trivial for the general public to misuse (copy-paste ready).
Mitigation Availability	Defences already exist and are deployable.	Mitigations are under development but not fully available.	No known mitigations or countermeasures.
Public Benefit of Openness	High: critical for reproducibility, security research, or public trust.	Moderate: some benefit but not urgent.	Low: little public interest beyond curiosity.
Recommendation Action	Full Release	Staged Release or Controlled Access (API, embargoed disclosure, safety monitoring)	Delay/Restrict Release until mitigations exist and risks are reduced.

OpenAI's decision to initially withhold GPT-2's 1.5B parameter model in 2019 was controversial but ultimately prescient. At the time, concerns centred on potential misuse for disinformation campaigns, spam generation and malicious automation. Table IV summarizes timeline, risk assessment process, stakeholder engagement, and mitigation steps, showing how gradual disclosure reduced perceived misuse risk over time.

Table IV: Case Study Analysis of OpenAI's Staged Release of GPT-2

Criterion	Assessment (2019)
Risk Severity	**Medium**: Capable of producing coherent synthetic news articles and spam.
Exploitability	**High**: Prompting and replication were trivial once weights were public.
Mitigation Availability	**Low**: Few detection or content provenance tools existed.
Public Benefit of Openness	**Medium**: Research benefits the real but not urgent public good.

i. Matrix Output (Staged Release):

OpenAI initially published the paper without full weights, releasing progressively larger models over time and monitoring misuse signals. As the community gained better detection methods and the observed misuse remained limited, OpenAI released the complete model six months later. This aligns with the matrix recommendation, i.e., a temporary restriction until mitigations improve.

Case 2: Vulnerability-Generating Outputs

Recent studies have shown that some LLMs can generate working exploit code for known vulnerabilities with minimal prompting. This capability can be valuable for penetration testers and defenders, but could also enable low-skill attackers to launch real-world exploits. Table V evaluates benefits against risks and notes whether governance measures were applied.

ii. Matrix Output (Controlled Access):

Developers should avoid open-weight release until mitigations (e.g., robust exploit filters, vulnerability reporting protocols) are in place. API-only access with monitoring can balance researcher needs with harm prevention. This case demonstrates the need for dynamic release strategies that evolve with the patching landscape.

Case 3: Open-Weight Model Releases (e.g., Llama 2)

Table V: Case Study Analysis of LLMs Producing Cyber Exploits

Criterion	Assessment
Risk Severity	**High**: Could enable cyberattacks at scale if exploited by malicious actors.
Exploitability	**High**: Outputs are often copied and pasted as executable.
Mitigation Availability	**Medium**: Some patches exist, but not for zero-day vulnerabilities.
Public Benefit of Openness	**Moderate**: Security researchers' benefit, but release could outpace defence capacity.

Meta's release of Llama 2 weights in 2023 sparked debate about whether unrestricted distribution of powerful models increases societal risk. Advocates highlighted the research benefits of open models, while critics warned about the potential for misuse. Table VI highlights the experiment setup, the types of vulnerabilities generated, adversarial prompting success rates, and mitigation strategies proposed by researchers to limit uncontrolled misuse.

iii. Matrix Output (Staged or Gated Release):

A proportional approach might have included gated access (researcher application process, usage tracking) or a phased release of progressively larger models. While Meta opted for full open release, the matrix would suggest balancing this decision with parallel

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

investment in provenance infrastructure and monitoring capacity.

Table VI: Case Study Comparison of Open-Weight Model Releases Such as LLaMA

Criterion	Assessment (2019)
Risk Severity	**Medium to High**: Capable of disinformation, deepfake text, and malicious code generation.
Exploitability	**High**: Anyone with commodity hardware can run the model locally.
Mitigation Availability	**Medium**: Some alignment layers exist but are easily removed.
Public Benefit of Openness	**High**: Significant value for independent safety auditing and democratization of research.

C. Advantages of the Matrix Approach

i. *Transparency:* Provides clear criteria, reducing the perception that release decisions are arbitrary or politically motivated.
ii. *Consistency:* Enables organizations to make comparable decisions across different model generations and capability thresholds.
iii. *Accountability:* Offers a principled justification that can be communicated to stakeholders, improving public trust.
iv. *Flexibility:* Allows for updates as technology evolves; the same matrix can be recalibrated for more powerful frontier models.

D. Limitations and Considerations

The matrix does not eliminate judgment calls. Risk severity is often uncertain, especially for novel capabilities where real-world misuse data are sparse. There is also a risk of false negatives (underestimating risk) and false positives (overestimating, leading to unnecessary restriction). To mitigate this, the matrix should be used by multidisciplinary committees including ethicists, domain experts, and security professionals. Furthermore, disclosure decisions should be revisited periodically; a model considered high-risk at launch may become safer to release as mitigations improve.

E. Beyond Individual Organizations

For maximal impact, this matrix should not be used in isolation. Industry consortia and policymakers could adopt it as a baseline standard, encouraging coordinated release strategies and reducing competitive pressure to release prematurely. Just as ISO/IEC 29147 standardized vulnerability disclosure, a shared AI disclosure framework could normalize proportional release as best practice. This would also facilitate cross-lab information sharing, helping prevent "race to the bottom" dynamics where the least cautious actor sets the de facto norm for everyone.

The Disclosure Decision Matrix provides a practical mechanism to translate moral reasoning into operational policy. By evaluating risk severity, exploitability, mitigation readiness, and public benefit, it guides developers toward proportionate disclosure choices. The case studies demonstrate that this approach is not only theoretically sound but also consistent with real-world decisions such as GPT-2's staged release. Institutionalizing such a matrix could help the AI community balance innovation and security in a principled, repeatable way.

V. POLICY RECOMMENDATIONS

The Disclosure Decision Matrix and Decision Grid presented above provide a principled foundation for individual model release decisions. However, their real value emerges when adopted as part of a broader ecosystem of governance. Model developers, policymakers, academic institutions, and civil society actors all play a role in shaping how disclosure norms evolve. In this section, we offer concrete policy recommendations designed to institutionalize proportional disclosure and reduce the risk of catastrophic misuse while preserving the benefits of open research.

A. Establish Multi-Stakeholder Risk Review Boards

One of the clearest lessons from biosecurity is the value of independent review. Institutional Biosafety Committees (IBCs) evaluate dual-use research of concern under confidentiality, providing a check on the judgment of individual laboratories. The AI field could adopt a similar model by creating AI Risk Review Boards at the institutional or industry level.

These boards should include AI researchers, ethicists, cybersecurity experts, and public interest representatives. Their mandate would be to review pre-publication risk assessments, using standardized tools such as the decision matrix, and to recommend one of three actions: (1) immediate release, (2) staged or controlled release, or (3) delayed release pending mitigations. By distributing responsibility, review boards reduce the moral burden on individual developers and lend legitimacy to decisions that might otherwise appear opaque or self-serving.

B. Normalize Staged Release Protocols

One of the key advantages of the matrix approach is that it supports graduated disclosure, not an all-or-nothing choice. Staged release protocols — such as releasing a technical paper first, then a smaller model, then larger models as risk is better understood — should become a standard practice for frontier models.

Organizations should document these release plans in advance and make them publicly available. This transparency helps preempt accusations of arbitrary gatekeeping and allows external researchers to prepare for upcoming releases. It also creates a feedback loop; if misuse signals appear during early stages, later releases can be paused or modified. Crucially, staged release should not be viewed as a permanent withholding of information but as a mechanism to align disclosure timing with the availability of mitigations.

C. Strengthen Model Access Controls and Monitoring

For cases where full release would be irresponsible, controlled access mechanisms are essential. API-based deployment, rate limiting, and user authentication can meaningfully reduce exploitability without blocking legitimate research. Logging and misuse detection systems should be built into access pipelines, enabling developers to identify when their models are being used to generate exploit code, phishing templates, or disinformation.

Such monitoring should respect user privacy and focus on aggregate misuse patterns rather than surveillance of

© Copyright: All rights reserved.

individual users. Where appropriate, developers should share anonymized misuse data with trusted third parties to improve the collective understanding of threat trends.

D. Invest in Mitigation Research and Open-Source Safety Tools

Delaying release is only ethically justifiable if it is paired with active efforts to close the risk gap. Model developers and governments should fund research into automatic exploit detection, content provenance systems, and biological hazard screening that can be deployed alongside powerful models.

Open-source safety tools are essential because they enable smaller labs and independent researchers to participate in responsible innovation. For example, if open-weight models are released with accompanying filters and detection systems, the risk of misuse is significantly lower than if those models are released without guardrails.

E. Create a Coordinated AI Vulnerability Reporting System

Just as cybersecurity has the CVE database and CERT coordination centers, AI needs a centralized vulnerability reporting mechanism. Researchers who discover dangerous capabilities or safety failures should have a trusted channel to report them without fear of legal or reputational reprisal.

This reporting system could maintain an embargoed vulnerability list, notifying major labs and infrastructure providers before disclosing to the public. Aligning with the matrix, disclosure timing would depend on the severity and availability of mitigation measures, ensuring that the public is informed when risks are manageable but not prematurely exposed to attack vectors.

F. Encourage Regulatory Backstops for High-Capability Models

While voluntary norms are a starting point, regulatory guardrails may be necessary for frontier models that pose systemic risks. Governments could require that models above a certain compute or capability threshold undergo risk assessment and review before open release. Such requirements should be narrowly tailored to avoid stifling innovation but strong enough to prevent the "race to release" that could amplify harm.

Importantly, regulation should be aligned internationally where possible, as AI research is global and unilateral restrictions may simply drive risky development elsewhere. International coordination bodies, potentially modelled on the International Atomic Energy Agency, could oversee compliance and facilitate information sharing about dangerous capabilities.

G. Foster a Culture of Ethical Responsibility

Finally, disclosure practices must be embedded in the culture of AI research. Ethical training for ML engineers and researchers should include case studies on dual-use dilemmas, much as medical education includes bioethics. Conferences and journals can require risk assessment statements alongside model releases, similar to conflict-of-interest disclosures.

Public communication also matters. When labs choose to delay release, they should clearly explain why, what criteria will trigger future release, and what mitigations are being pursued. This transparency builds trust and avoids public backlash based on perceptions of secrecy or corporate control.

Policy recommendations grounded in the decision matrix go beyond abstract moral reasoning; they provide a concrete blueprint for building an ecosystem of responsible disclosure. Multi-stakeholder review boards distribute decision-making authority, staged release protocols align openness with readiness, and vulnerability reporting systems provide early warning. Together with investment in mitigations and regulatory coordination, these measures can ensure that robust generative AI systems are deployed in a way that maximizes public benefit while minimizing catastrophic risk. The final section synthesizes these insights and highlights directions for future research.

VI. CONCLUSION

Generative AI has transformed the landscape of knowledge production, but its power comes with profound ethical responsibilities. This paper has argued that disclosure decisions about powerful large language models cannot be treated as an afterthought or a purely pragmatic consideration. By analyzing the problem through utilitarian, deontological, and virtue-ethical lenses, we demonstrated that neither radical openness nor blanket secrecy is morally defensible. Instead, the ethically justified approach is one of proportional disclosure, where release decisions scale with risk severity, exploitability, and the availability of mitigations.

Our principal contribution is the Disclosure Decision Matrix, a practical tool that operationalizes moral reasoning into actionable criteria. By evaluating risk severity, exploitability, mitigation readiness, and the public benefit of openness, the matrix guides developers toward decisions that are transparent, consistent, and ethically grounded. Through case studies of GPT-2's staged release, vulnerability-generating outputs, and open-weight model releases, we illustrated how this framework can be applied in real-world scenarios.

Policy recommendations built on this framework emphasize the need for institutionalized review processes, staged release protocols, controlled access mechanisms, and shared vulnerability reporting systems. These measures collectively reduce the risk of catastrophic misuse while preserving the benefits of open research and democratic access to AI technology.

Future research should focus on refining quantitative methods for risk assessment, including more robust models for estimating the probability and impact of misuse. Cross-disciplinary collaboration with cybersecurity, biosecurity, and risk analysis experts will be critical to building predictive frameworks that can inform disclosure timing. Additionally, as models continue to scale, international coordination will become essential to prevent unilateral release decisions from undermining global safety efforts.

The task ahead is to transform these recommendations into widely adopted norms and, where necessary, regulatory standards. By embedding ethical reasoning into the heart of AI release governance, the research and policy

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

community can ensure that generative AI develops not only as a powerful technology but as a responsible one, aligned with the broader public good.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/Competing Interests:** Based on my understanding, this article does not have any conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is crucial for affirming its impartiality, as it was conducted without any external influence.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The data and materials supporting the findings of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. N. Carlini et al., "Emergent risks in large language models," arXiv preprint arXiv:2304.15004, 2023. Available on: https://arxiv.org/abs/2304.15004.
2. M. Urbina et al., "Dual-use concerns in synthetic biology," Frontiers in Bioengineering and Biotechnology, vol. 8, 2020.
3. B. Nosek et al., "Promoting an open research culture," Science, vol. 348, no. 6242, pp. 1422–1425, 2015, DOI: https://doi.org/10.1126/science.aab2374
4. OpenAI, "GPT-2: 1.5B Release Notes," OpenAI Blog, Nov. 2019. Online: https://openai.com/index/gpt-2-1-5b-release.
5. Meta AI, "Introducing Llama 2," Meta AI Blog, July 2023. Online: https://about.fb.com/news/2023/07/llama-2.
6. Walshe, T. and Simpson, A., "Your Vulnerability Disclosure is Important to Us," Computers & Security, 2022, DOI: https://doi.org/10.1016/j.cose.2022.102895
7. Brundage, M. et al., "The Malicious Use of Artificial Intelligence," arXiv, 2018, DOI: https://doi.org/10.48550/arXiv.1802.07228
8. Carlini, N. et al., "Emergent Risks in Large Language Models," arXiv, 2023, DOI: https://doi.org/10.48550/arXiv.2304.15004
9. Weidinger, L. et al., "Taxonomy of Risks Posed by Language Models," arXiv, 2022, DOI: https://doi.org/10.48550/arXiv.2112.04359
10. A. F. Martinho, T. Paiva, and M. J. Ribeiro, "The risks of voice cloning and deepfake audio: A review of recent cases and regulatory responses," *Computers & Security*, vol. 135, p. 103521, 2024. DOI: https://doi.org/10.1016/j.cose.2024.103521
11. M. Schick and S. Vogl, "AI voice impersonation and telecom regulation: Emerging challenges for fraud prevention," *Telecommunications Policy*, vol. 48, no. 2, p. 102634, 2024. DOI: https://doi.org/10.1016/j.telpol.2023.102634
12. E. Zou, L. Song, and R. Shokri, "Universal and transferable jailbreaks on aligned language models," *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pp. 1234–1248, 2024. DOI: https://doi.org/10.1109/SP46215.2024.00093
13. Biderman, S. et al., "Lessons from the Trenches on Deploying Open Language Models," arXiv, 2023, DOI: https://doi.org/10.48550/arXiv.2302.13971
14. Bommasani, R. et al., "Foundation Models in AI," arXiv, 2022, DOI: https://doi.org/10.48550/arXiv.2207.05221
15. Shevlane, T. and Dafoe, A., "The Offence-Defence Balance of Scientific Knowledge," AIES '20, DOI: https://doi.org/10.1145/3375627.3375815
16. Solaiman, I. et al., "Release Strategies and the Social Impacts of Language Models," arXiv, 2019, DOI: https://doi.org/10.48550/arXiv.1908.09203
17. Hansson, S.O., "How Extreme is the Precautionary Principle?" Sci Eng Ethics, 2017, DOI: https://doi.org/10.1007/s11948-017-9900-y

AUTHOR'S PROFILE

Fahd Malik is an accomplished AI and Data Technical Manager with over eight years of experience driving digital transformation across the telecommunications sector. He specializes in delivering data-driven and AI-powered solutions, managing complex BSS, CRM, and automation initiatives from strategy to execution. At Zain KSA, he has led projects that optimized customer engagement, reduced operational costs, and enhanced service efficiency through advanced analytics and AI integration. Fahad holds certifications in PMP, ITIL, Lean Six Sigma, and AWS, and is pursuing a Master's in Digital Transformation at IE Business School. His interests span AI, blockchain, metaverse technologies, and cloud innovation.

Muhammad Raza ul Haq is a seasoned telecommunications professional with over 20 years of experience in Business Support Systems (BSS), covering domains such as billing, CRM, mediation, and revenue assurance. He has worked across Pakistan, Bahrain, and Saudi Arabia, contributing to major telecom implementations and post-migration support. His research focuses on emerging technologies in the telecom sector, including 5G, Artificial Intelligence (AI), Machine Learning (ML), and Large Language Models (LLMs). He explores their role in enhancing digital transformation, automation, and customer experience.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.