The Open-Weight Dilemma

Cyber Policy When Control on AI is Lost

Alfonso De Gregorio

PWNSHOW

The Open-Weight Dilemma

Cyber Policy When Control on AI is Lost

by

Alfonso De Gregorio

PWNSHOW

Summary

Open-weight general-purpose AI models—whose trained weights are publicly released for anyone to use or modify—offer significant benefits in innovation and transparency. However, they also introduce substantial cybersecurity risks, as demonstrated by models like DeepSeek-R1 achieving over 90% success on offensive cyber knowledge tests [6] . By enabling a broader range of actors to automate and scale cyberattacks, these openly released large language models (LLMs) challenge traditional defense paradigms and regulatory approaches. This article analyzes specific threats amplified by open-weight LLMs—including accelerated malware development, sophisticated phishing and social engineering, and faster discovery of software vulnerabilities. It critically assesses current regulatory frameworks, notably the EU Artificial Intelligence Act (AI Act) and the draft General-Purpose AI (GPAI) Code of Practice, identifying major gaps that arise from the loss of control once model weights are widely available [7]. In particular, the AI Act's provisions (and exemptions) were largely designed for closed or controlled AI systems, and thus exhibit blind spots when applied to openly distributed models [4]. Standard security mitigations (like usage monitoring or access gating) become ineffective once a model is released to the public. Drawing on the author's consultation input to the European Commission, we incorporate expert clarifications on how the AI Act should be interpreted for open models—for example, treating the act of fine-tuning to remove safety features as a "substantial modification" that shifts accountability to the modifier [3, 2]. We propose a path forward focusing on pragmatic policy adaptations and technical measures: evaluating and controlling specific high-risk capabilities rather than banning entire models; shifting security obligations to the actors with actual control (e.g. downstream deployers) at each stage; promoting defensive AI innovations to counter malicious use of AI; and fostering international collaboration on standards and cyber threat intelligence sharing. These recommendations aim to ensure security and accountability in the age of open-weight LLMs without unduly stifling open innovation.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
API	Application Programming Interface
EU	European Union
FLOPS	FLoating-point Operations Per Second
GPAI	General Purpose AI
LLM	Large Language Model
OECD	Organisation for Economic Co-operation and Development
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OSI	Open-Source Initiative
RLHF	Reinforcement Learning from Human Feedback
TACTL	Threat Actor Competency Test for LLMs
TTP	Tactics, Techniques, Procedures

1

Introduction

Motivation: The rapid progress of general-purpose AI (GPAI), especially large language models (LLMs), has led to a proliferation of open-weight models – those whose parameter weights are made public for anyone to download and use. This openness has spurred a "Cambrian explosion" of innovation by enabling researchers and smaller companies to build on state-of-the-art AI without prohibitive cost [5]. Yet, the very openness that democratizes AI development also heightens concerns about misuse, particularly in cybersecurity domains [9]. Advanced AI models have demonstrated alarming offensive cyber capabilities. For instance, evaluations by MITRE's OCCULT framework found that an open-release model, DeepSeek-R1, could correctly answer over 90% of challenging questions in a threat actor tactics benchmark [6] – essentially achieving expert-level knowledge in cyber offensives. The release of DeepSeek-R1 (a 685-billion-parameter reasoning model developed openly in China) has galvanized policymakers by highlighting how unexpected and powerful capabilities can emerge from open models [10]. Additionally, specialized malicious AI tools are appearing on the underground: Xanthorox AI, a self-contained AI-driven hacking platform surfaced on darknet forums in early 2025 [8]. Unlike benign chatbots jailbroken for misuse, Xanthorox was purpose-built by attackers, offering an all-in-one system for generating malware code, exploiting vulnerabilities, and orchestrating autonomous cyber attacks [1]. These developments signal a paradigm shift—potent AI capabilities once limitedtowell-resourced state actors are now accessible to criminal syndicates, hacktivists, and lone hackers.

Problem Statement: Open-weight LLMs threaten to lower the barrier to entry for sophisticated cyberattacks. They can automate phishing campaigns by generating convincing spear-phishing emails at scale, rapidly produce polymorphic malware or ransomware code, and even assist in discovering zero-day vulnerabilities by analyzing software or configurations [4]. Such capabilities, if broadly available, could overwhelm current cybersecurity defenses which are already strained. Traditional defensive strategies (e.g. signature-based malware detection, manual threat hunting) may be ill-equipped to handle AI-accelerated attacks that adapt and multiply faster than humans can respond. Moreover, once model weights are released publicly, the model's original creators lose visibility and control over its use. This creates a *"mitigation gap"*: security measures that rely on keeping a model behind an API or monitoring its outputs cannot be enforced when anyone can run a copy locally. For example, cloud-based AI services can throttle or filter potentially harmful outputs, but an openly downloadable model can be modified to remove those safety filters entirely. The loss of control inherent in open distribution thus renders many standard AI governance mechanisms ineffective [4]. This gap poses a serious challenge for policymakers and practitioners alike – how do we reap the benefits of open AI innovation while managing the new cyber risks?

Research Questions: This article addresses several key questions: (1) What specific offensive cyber capabilities are enhanced or enabled by open-weight LLMs? (2) How do current regulatory frameworks, such as the EU AI Act and associated policies, address (or fail to address) the risks posed by publicly available AI model weights? (3) Why do traditional AI risk mitigation strategies break down once model weights are openly released? and (4) What policy and technical interven-

tions could effectively mitigate the cybersecurity risks of open-weight AI, while preserving the advantages of openness? We tackle these questions through an interdisciplinary analysis, incorporating both technical evidence (from recent AI evaluations and real-world incidents) and policy analysis (of existing regulations and expert consultation feedback).

Structure: The remainder of this article is organized as follows. Chapter 2 provides background on offensive cyber operations and defines the open-weight model phenomenon, distinguishing it from traditional open-source software. We also review related work on AI security. Chapter 3 examines the cybersecurity threats amplified by open-weight LLMs, with concrete examples of how such models can be misused for cyber offense. Chapter 4 analyzes the current regulatory landscape—focusing on the EU AI Act and the draft GPAI Code of Practice—and identifies gaps or mismatches when these rules are applied to open models. We integrate clarifications from the EU's AI Act consultation to highlight specific ambiguities (like the "open-source exemption" and responsibilities for model modifiers). Chapter 5 discusses mitigation approaches, both technical and cooperative: leveraging "AI to fight AI" in cybersecurity, and the importance of international collaboration (standards, threat intelligence sharing, and norms) given the global availability of these models. Chapter 6 offers concrete policy and technical recommendations, such as targeting regulation at high-risk capabilities rather than entire models, introducing controlled release mechanisms and transparency obligations, incentivizing responsible open model release practices, and clarifying legal definitions to ensure accountability falls on the appropriate actors. Chapter 7 concludes with key takeaways and future directions, arguing for a balanced approach that secures open-weight AI without stifling its innovative potential.

2

Background

2.1. Offensive Cyber Operations and AI's Emerging Role

For decades, sophisticated cyberattacks (such as state-sponsored espionage, intelligence gathering, sabotage, or complex malware campaigns) were largely confined to well-resourced actors with specialized expertise. In recent years, however, the landscape has "democratized" to an extent—underground markets offer cybercrime-as-a-service, from ready-made phishing kits to ransomware toolkits, lowering the skill needed to launch attacks. The advent of powerful AI systems stands to accelerate this trend. Generative AI can produce human-like text, code, or even multimedia, enabling threat actors to automate tasks that previously required human creativity or effort. For example, an AI model can generate highly persuasive phishing emails tailored to a target, or mutate malware code to evade antivirus detection, in a fraction of the time a human adversary would need.

Early closed AI systems (like OpenAI's Codex or ChatGPT) hinted at these dual-use possibilities, but they were accessible only via restricted APIs with content filters. The real shift came with the release of open- weight models that anyone can run offline. These include large language models such as Meta's LLaMA and its derivatives, and newer entrants like Mistral and DeepSeek families. When Meta released LLaMA's weights openly in 2023, it marked a turning point—unleashing a wave of community- driven development, but also raising the specter of misuse. Unlike traditional open-source software (where source code is published, but running the software still requires skill and context), open-weight AI puts a ready-to-use capability directly into users' hands. A sufficiently advanced LLM can be instructed (or fine-tuned) to perform harmful tasks (e.g. generating malware or disinformation) unless explicitly constrained.

Defining "Open-Weight" vs. "Open-Source": It is important to clarify terminology. Open-weight models refer to AI models whose learned parameters (weights) are publicly released, allowing anyone to download and instantiate the model. This is a looser concept than open-source AI as defined by the Open Source Initiative (OSI). The strict OSI definition would require not just weights, but also full access to training data and training code under an open license. In practice, most "open" AI models today are open-weight but not fully open-data—for instance, LLaMA or Mistral released model weights but used proprietary or filtered training datasets. Thus, while these models are widely accessible, they might not meet the purist open-source criteria. This distinction becomes relevant in policy (as we'll see with the EU AI Act's exemption for "free and open-source" AI). Nonetheless, the key point is that open-weight distribution transfers the model's capabilities to the public domain. Developers worldwide can run and fine-tune these models on their own infrastructure without oversight from the original provider.

2.2. AI Capability Evaluations Highlight Offensive Potential

Recent research evaluations have empirically demonstrated the offensive potential of large LLMs. MITRE's OCCULT framework is a notable example. OCCULT introduced benchmarks like a Threat Actor Competency Test for LLMs (TACTL) to gauge how well an AI can perform tasks or answer

questions that a cyber attacker might encounter [6]. The startling finding was that for the first time, an AI model (DeepSeek-R1) scored above 90% on these adversarial knowledge tests, outperforming many prior models [6]. DeepSeek-R1 is an especially relevant case: developed by an AI lab and released as an open-weight model, it possesses a combination of reasoning and coding abilities optimized for solving complex problems. Its high OCCULT scores mean it can recall or generate solutions to hacking challenges (e.g. how to exploit a certain vulnerability or bypass a security control) at near-expert accuracy [6]. While this does not automatically make DeepSeek a fully autonomous "hacking AI", it dramatically lowers the expertise required to plan and execute advanced techniques—the model can supply knowledge and even step-by-step guidance.

Another source of insight comes from real-world threat activity. The appearance of Xanthorox AI on darknet forums (March 2025) illustrates how malicious actors are actively integrating AI into their toolchain. Xanthorox is described as a modular, autonomous attack platform offering various AI-driven modules for tasks like vulnerability scanning, exploit generation, and even voice- or image-based social engineering [7]. Notably, it does not rely on external APIs or known models (like trying to prompt ChatGPT); instead it carries custom-trained local models, specifically to avoid any oversight or usage restrictions. By being hosted on the attackers' own servers with no external dependencies, such a platform evades the "kill-switch" that companies or cloud providers might invoke if an AI is being misused. This indicates a strategy by cybercriminals to stay independent and under the radar with their AI tools [7]. The developers of Xanthorox tout features like handling code generation and vulnerability exploitation in a stealthy manner, and being adaptable via its modular design. In essence, this is an offensive AI ecosystem emerging in parallel to the legitimate AI ecosystem.

These examples underscore that AI's offensive capabilities are no longer theoretical. They are here now, and open-weight releases act as an accelerant. An analogy can be made to the proliferation of strong encryption in the 1990s: once the algorithms were public, control shifted irreversibly. Similarly, with advanced AI models out in the wild, the focus must shift from trying to contain capabilities to instead adapting defenses and policies to a new reality.

3

Cybersecurity Threats Amplified by Open-Weight LLMs

Open-weight LLMs can enhance or enable several offensive cyber techniques. Below we highlight some key threat vectors that are magnified by the availability of such models:

- *Automated Social Engineering:* LLMs excel at generating fluent, contextually relevant text. Malicious actors can leverage an open model to craft highly convincing phishing emails, fraudulent messages, or even real-time chat responses for scams. Unlike template-based phishing, AI-generated lures can be tailored to the target (scraping their social media or organization info) and varied infinitely to evade spam filters. Deepfakes in text form – such as impersonation of a colleague's writing style – become trivial. Open models can also produce multilingual phishing at scale, broadening the attacker's reach. The result is more effective social engineering campaigns that can be launched by relatively low-skilled attackers, since the model provides the "brains" of the con.
- *Malware Development and Evasion:* Generative models can write code, including malware code. With an open-weight model, an attacker can interactively develop malicious software, ask the model to refine or obfuscate it, and even generate polymorphic variants. This can accelerate the creation of new malware strains. Furthermore, models can be instructed to find and exploit vulnerabilities in code. For instance, given a snippet of software, an AI might identify a buffer overflow and then help generate exploit code for it [4]. In evaluations, LLMs have demonstrated the ability to produce working exploits for known vulnerabilities when prompted with the right context (though often these models were aligned to refuse if asked overtly – a restriction that can be removed in an open version). Attackers can also use AI to generate evasive malware that morphs its indicators (like function names or strings) to defeat signature-based detection. Each of these tasks – once requiring dedicated expertise – can be partially or fully automated with a competent open model.
- *Scaling of Spear-Phishing and Disinformation:* Beyond one-off phishing emails, open models can generate entire social media influence campaigns: fake personas with consistent backstories, posts, and interactions, all orchestrated by AI. They can produce tailored scam communications on a massive scale (e.g. personalized scam chats to thousands of targets on messaging platforms). They can also be used to automate reconnaissance: writing code to crawl and summarize targets' online presence, then formulating social engineering schemes. The barrier to conducting broad phishing or disinformation operations is thus much lower – one needs fewer human "workers" since the AI does the content creation. This could lead to an overall increase in volume of attacks and potentially the emergence of fully autonomous phishing bots.
- *Faster Vulnerability Discovery:* AI models can assist in finding vulnerabilities by analyzing source code or configurations much faster than humans. A model fine-tuned on code analysis could flag potential security bugs or dangerous patterns within large codebases in minutes, a task

that might take a human auditor days. Open-weight models allow attackers to perform such analysis on their own hardware, without needing to send data to an API (which might be monitored). Early research suggests that with appropriate training, AI can uncover novel vulnerabilities (including those not previously public) by pattern-matching against its training knowledge and known exploit techniques. Attackers armed with such tools could dramatically shorten the zero-day discovery cycle. Additionally, once a vulnerability is known, AI can help optimize and test exploits against it, as mentioned earlier.

- *Offensive AI Agents:* Perhaps the most concerning long-term scenario is the creation of semi-autonomous AI agents that can carry out multi-step cyber operations. For example, an AI agent could be instructed to achieve a goal ("exfiltrate sensitive data from target X"). It could then plan an operation: reconnoiter the target (using AI vision or NLP for data analysis), identify an entry point, craft an exploit or phishing email, execute it, escalate privileges, and extract data – adjusting its strategy based on success/failure at each step. While this composite capability is nascent, pieces of it are being developed. Open-weight models are an essential ingredient because such agents must run continuously and cannot rely on a third-party API that might block malicious actions. A publicly released model can be integrated and internally "prompted" by agent software without constraints. Projects like AutoGPT and others have shown AI agents chaining tasks autonomously; in the wrong hands, similar architectures could be directed towards cyberattacks. Xanthorox AI's design (multiple specialized models coordinating tasks) is an early example of this modular attack agent approach [*7*].

Implications for Defense: These amplified threats mean that many traditional cybersecurity measures could be overwhelmed. Human analysts and conventional security tools might struggle to cope with AI-augmented attack speed and volume. For instance, if phishing attacks multiply via automation, even a 99.9% effective filter would still let more incidents through simply due to scale. If malware iterates its code faster than signatures can be written, detection lags behind. The use of AI also means attacks may present novel patterns that evade detection systems trained on historical (non-AI) data. In summary, open-weight AI models act as force multipliers for attackers. They provide asymmetric advantages: a single malicious actor with a powerful model can do damage far beyond what they could manage alone. This raises the stakes for cybersecurity—requiring not just incremental improvements, but a rethinking of defense and policy, as we explore in the following sections.

4

Regulatory Setting and Policy Gaps

To address these emerging risks, we must examine the current regulatory frameworks governing AI and see how they apply to open-weight models. The European Union's AI Act is a landmark attempt to regulate AI, and it explicitly covers general-purpose AI models (GPAI) which include large language models. Alongside, the European Commission is developing a GPAI Code of Practice, a voluntary set of commitments for AI providers to implement ahead of the Act's enforcement. We analyze these frameworks with a focus on cybersecurity and open-weight issues. We also incorporate insights from expert consultations (including the author's input to the EU) to identify ambiguities and suggest clarifications.

4.1. The EU AI Act: Overview and Applicability to Open Models

The EU AI Act is a risk-based regulation categorizing AI systems into tiers of risk (unacceptable risk, high-risk, limited, minimal) with corresponding requirements [3]. Many provisions target high-risk AI systems—those used in sensitive areas like employment, biometric identification, critical infrastructure, etc. Initially, general-purpose AI like GPT-type models were not clearly within scope, but later drafts introduced provisions for GPAI, recognizing that such models can be integrated into high-risk applications by others. The Act has extraterritorial reach: any model or system that affects people in the EU could fall under its rules, even if developed elsewhere [4].

High-Risk Obligations: For an AI system classified as high-risk (either by its inherent function or its deployment in a high-risk domain), the Act imposes a set of obligations on the provider of that system [5]. These include: establishing a risk management system (continuous assessment and mitigation of risks, per Article 9), ensuring high-quality training data (Article 10), keeping technical documentation and logs (Articles 12 and 13) for traceability, ensuring transparency to users (e.g. disclosures when AI is used, Article 13), implementing human oversight (Article 14), and ensuring robustness, accuracy and cybersecurity (Article 15), among others [5]. Non-compliance can lead to hefty fines.

A fundamental assumption in these obligations is that the provider—the entity putting the system on the market—has control over the system's development and can implement these measures. This assumption holds for traditional software or closed AI services, where one company builds and deploys the system (and can update it, monitor it, etc.). However, with an open-weight model, this assumption breaks down. The provider (e.g., the company that initially trained and released the model) does not control what happens after release: they cannot track who is using the model or how, cannot directly monitor outputs or updates, and often cannot even determine if the model is being used in a high-risk context.

Open-Source Exemption: The EU AI Act, in its latest draft form, contains an exemption for "free and open-source AI software". Specifically, it generally exempts from its requirements those AI components that are made available under open-source licenses, with certain conditions [4]. The rationale is to avoid stifling open research and hobbyist innovation. However, this exemption is raising considerable debate and confusion. The crux is what counts as "open-source" in this context. If one takes a strict interpretation, it might mean that only AI projects releasing both model and

training data openly (truly open-source in OSI terms) are exempt [4]. Under that view, many open-weight models (like LLaMA, Mistral, DeepSeek) would not qualify as they did not open their training datasets [4]. They would then technically be subject to the Act's obligations as GPAI providers—obligations which, as noted, they realistically cannot fulfill post-release. This would put these open model providers in an untenable position: facing compliance duties (record-keeping, post-market monitoring, etc.) that are practically impossible once the model is public, potentially dissuading them from releasing models at all [4].

On the other hand, a looser interpretation might treat "open-weight" distribution itself as sufficient to claim the open-source exemption (i.e., if weights are released for free use, consider that open-source even if the data isn't public) [4]. That would mean models like LLaMA and DeepSeek-R1 are exempt from most obligations, even though they pose significant risks. This approach might undermine the Act's intent by letting highly capable models slip through oversight merely because they are open. Indeed, a very powerful model could be completely unregulated if it is deemed open-source, which is paradoxical if our concern is misuse. The European Commission's draft GPAI Code of Practice has not clarified this ambiguity either [4].

Current signals suggest regulators are aware of this dilemma. In fact, the Act's text was adjusted to exclude "open source put on the market" from many requirements unless it's part of a final product. Yet, the precise scope remains murky. The author's consultation feedback highlighted this as "the central point of ambiguity"—if interpreted strictly, innovation is stifled; if loosely, risks are unmitigated [5]. There is consensus that guidance is needed to thread the needle: ensure that bona fide open research isn't crushed by compliance burdens, while not creating a loophole for dangerous open models to be entirely ungoverned [5].

Substantial Modification and Provider Liability: The AI Act defines a "provider" as not only the original developer, but also anyone who substantially modifies an AI system or repurposes it (Article 25) [5]. This is highly relevant for open-weight models because it means if a third party takes an open model and significantly changes it (e.g. fine-tunes it for a high-risk use), that third party could legally become the "provider" of a new high-risk system, with all associated obligations [5]. The Act lists examples of substantial modifications: changing the intended purpose, or making a modification that affects compliance. However, it has been noted that it's unclear whether removing safety constraints from a model counts. The author's input strongly argued that fine-tuning an open model to bypass or remove its safety features must be explicitly deemed a "substantial modification." [5, 3]. For instance, if a company releases an LLM with guardrails (it refuses to output violent or illicit content), and an adversary downloads the weights and fine-tunes the model to disable those guardrails (sometimes termed "alignment removal" or model "unfiltering"), this should unequivocally count as a substantial change. Why? Because the model's risk profile is now drastically altered – it could produce harmful content or instructions freely. If the Act recognizes this as a substantial modification, then the modifier assumes the role of provider and is subject to the law (while the original provider is off the hook for those modifications) [5, 2]. This clarification would close a potential loophole where an open model's creator might otherwise wrongly be held responsible for misdeeds committed with a modified version that they have no control over [5]. It also ensures there is accountability – the bad actor can't claim innocence under the cover of using an open model.

Crucially, Article 25 of the Act already intends this transfer of obligations, but the consultation feedback emphasizes making it crystal clear for the open-weight scenario. Without it, we risk a situation where malicious actors exploit open models to create dangerous derivatives while responsibility remains ambiguous [5]. Encouragingly,inthelatest drafts of the AI Act and the Code of Practice, there are signs of embracing this principle. Recital 109 of the Act (reflected in the Code) recognizes that when a model is modified or fine-tuned, the original provider's obligations should be limited to their part, and the modifier should take on new obligations [2]. This proportional approach is essential to avoid punishing open model creators for things beyond their control, and to square the circle of openness vs. accountability.

Obligations Impossible to Enforce Post-Release: Another major gap is that certain obligations simply cannot be fulfilled once a model is open. For example, Article 12 (Record-keeping) expects providers to log the AI system's operation to some extent. A cloud AI service can do this (log queries, monitor outputs), but an open-weight model running on someone's PC generates no records accessible to the original provider. Similarly, Article 13 (Transparency) might require that users are informed the content is AI-generated or that the system can be identified (watermarking outputs,

etc.). However, as pointed out in consultation, watermarking can often be removed by a downstream user [5]. If someone has full model weights, they could tweak or fine-tune the model to eliminate any built-in watermarks or tracking, or simply not use the watermarked decoder. Thus, obligations that assume the provider can implement persistent measures in the model or monitor its use become unenforceable in open release scenarios [5, 4].
The author's consultation comments recommended that guidelines acknowledge these technical realities. Specifically, regulators should focus on what obligations can feasibly be met pre-release, rather than imposing post-release duties that are moot [5]. For instance, instead of requiring an open model provider to somehow monitor all uses (impossible), the emphasis should be on thorough documentation at release: model cards detailing capabilities/limitations, disclosure of known risks, results of red-team or safety testing, etc. We will discuss documentation expectations shortly. The point here is that a shift in mindset is needed – from continuous control to one-time responsible release. Once the "genie is out of the bottle", the focus of regulation should move downstream.

High-Risk Use by Downstream Actors: A general-purpose model per se might not be high-risk, but if someone uses it in a high-risk application (say, a medical diagnosis tool or a system for CV-scanning in hiring), then that application is high-risk. The AI Act envisions that in such cases the downstream deployer has obligations to ensure the system complies. But with open models, the downstream actor could be anyone, anywhere – not necessarily known to or in contract with the model provider. Article 25(4) of the Act tries to address this by saying basically that various parties in the value chain (importers, distributors, etc.) should cooperate via written agreements to ensure compliance, including passing information and technical assistance from model providers to downstream system providers [5]. This again presumes an ecosystem of contractual relationships. With an open model published on the internet, there's no direct relationship or contract between the original provider and a random developer who downloads it [5]. Therefore, those Article 25(4) cooperation mechanisms break down. You cannot have a written agreement with anonymous public downloaders; you often don't even know who they are.
The Act does carve out that third parties who just provide tools under open-source licenses are exempt from those value chain obligations, instead encouraging voluntary documentation practices (model cards, etc.) [5]. This is good in that it acknowledges the impracticality. However,it raises the question: how do we ensure downstream users still get the necessary information to use the model safely in a high-risk context? The answer must be: through public documentation and transparency provided at model release, since formal agreements are infeasible [5].
Indeed, one of the author's recommendations is that a foundational model provider releasing an open-weight model should treat the act of release almost like a one-time handoff of as much relevant information as possible – a "dowry" of documentation that enables any downstream user to understand the model's intended uses, limitations, risks, and how to mitigate them [5]. This would include: detailed Model Cards (describing what the model should and shouldn't be used for, and its performance characteristics) [5]; Dataset transparency (summaries of what data went into training, including whether any sensitive or domain-specific data like exploit code was included, which might give the model particular strengths) [5]; Capability and Safety Evaluations (sharing results from internal red-teaming or benchmarks such as OCCULT to flag the model's ability to produce offensive outputs – effectively a Responsible AI labeling of known dangerous capabilities) [5]; documentation of any embedded safety mechanisms (if the model was aligned via fine-tuning or RLHF, what it can or cannot do, and known ways those could be bypassed) [5]; and technical specs needed for safe deployment (like architecture, compatibility, etc.) [5]. By providing this upfront, the model developer assists any future high-risk deployer in meeting their obligations (risk assessment, transparency, etc.), even though they can't directly assist them later [5]. In essence, the responsibility shifts: the upstream provider's job is to inform and warn, whereas the downstream user's job is to implement and uphold appropriate safeguards for their specific use.

Summary of Gaps: Overall, the EU AI Act and similar regulations were not originally conceived with open, uncontrollable distribution in mind. As a result, they exhibit significant blind spots regarding open-weight models [4]. Key gaps include: ambiguity in the open-source exemption; the mismatch of accountability (the provider vs. modifier issue) in open contexts; obligations that assume continuous control which is absent for open models; and the lack of explicit guidance on how to handle the "mitigation gap" after release [5]. Without adaptations, there is a risk either that (a) open model developers will be deterred or face legal uncertainty, or (b) that open models become

a regulatory loophole exploited by malicious actors. Neither outcome is desirable.

4.2. The GPAI Code of Practice and Other Policy Initiatives

The General-Purpose AI Code of Practice is an EU-led voluntary initiative inviting AI providers to commit to certain best practices in the interim before the AI Act is enforceable. Drafts of the Code (as of early 2025) cover areas like transparency, risk management, and safety, mirroring many AI Act principles but in a non-binding format. The Code has seen multiple iterations (three drafts as of March 2025) and involves industry input. It is significant because major AI firms may sign on, and it can serve as a blueprint for compliance. From what we glean in the drafts, the Code of Practice emphasizes documentation (model documentation forms, transparency reports) and making information available to regulators upon request [2]. It also touches on the scenario of model modification: it defers to the AI Office to clarify when a modified model counts as new (essentially the fine-tune as new model question) [2]. Notably, the Code explicitly acknowledges that an open-source model with no usage restrictions can't really have an "Acceptable Use Policy" – so it provides guidance on how to document intended use, or even mark it as not applicable if the license is completely open [2]. This is a pragmatic nod to the reality of open models. However, the Code does not fully resolve the open-weight issues. The draft doesn't yet clarify the open-source exemption nuance we discussed (strict vs loose) [4]. It largely leaves it to the forthcoming AI Office to handle definitions and edge cases. This isn't surprising, as the Code is more about voluntary commitments (like "we will document X, we will assess Y"). One encouraging part is in the Transparency Recitals of the Code: "Signatories recognise that in case of a modification or fine-tuning of a model, the obligations for providers should be limited to that modification or fine-tuning" [2], reinforcing that proportional allocation of responsibility. So the Code's ethos supports our earlier points on shifting obligations to the party with control. It also stresses the need for downstream providers to get information from upstream (Recital 101 and others) [2].

Outside the EU, other policy discussions are relevant. For example, in the United States, there's an ongoing debate about imposing liability for AI. The reference to California's SB 1047 (a proposal in 2024) in the earlier text [4] is illustrative: that bill considered requiring "reasonable care" by AI deployers to prevent harm, and potentially broad liability if AI causes damage. Such broad-strokes regulation, if applied to open models, would likely place an impossible burden on open providers – as they cannot reasonably prevent misuse once the model is out [4]. It might force them to refrain from release due to fear of liability for others' actions. This again implicitly favors closed AI (where a company can maintain more control and demonstrate due care). We mention this to highlight a trend: without careful nuance, regulations aimed at AI safety could unintentionally choke off open collaboration and push everything behind corporate or governmental walls, which may not actually be the safest outcome either (security through obscurity has its own risks).

Another relevant global effort is the development of AI risk management frameworks (e.g., NIST's AI Risk Management Framework in the US) and sector-specific guidelines for AI security (like proposals at the OECD or United Nations level for responsible AI in military, etc.). Most of these are high-level and do not yet tackle the specific scenario of open-source AI aiding cybercrime. However, they contribute to an evolving normative landscape. For instance, if international norms declared that developing or distributing AI for malicious purposes is unacceptable, that could bolster efforts to stigmatize platforms like Xanthorox. But norms are voluntary and violators like cybercriminals are unlikely to be swayed.

In summary, today's policy toolkit is underprepared for open-weight AI's challenges. The EU AI Act is the most concrete approach on the horizon. It will likely need interpretive guidance to apply effectively to open models (some of which can be given via the AI Office or the Code of Practice). Key clarifications needed include: how to treat open model releases under the law (perhaps carving out a special category of GPAI with certain partial obligations), how to enforce documentation in lieu of control, and how to shift accountability to where it belongs in the value chain. The notion of focusing obligations on theparty "with actual control at each stage of the AI lifecycle "has emerged as a guiding principle [5]. That means, for example, the foundational model provider is responsible for transparency and addressing foreseeable risks in the model at release, whereas an entity that later fine-tunes or deploys it for a specific high-risk use is responsible for that specific use (conducting a new risk assessment, ensuring the end-system's compliance, etc.) [5]. This aligns with

general product liability concepts (e.g., a car manufacturer isn't liable if someone later installs illegal modifications causing harm – the modifier is).

The challenge is implementing this in practice, given the anonymity and global nature of open-source. It may require new mechanisms to identify when a model has been substantially modified and by whom – perhaps something like model "passports" or hashes that can indicate provenance. It also raises enforcement questions: how to hold a rogue fine-tuner accountable if they operate in a non-cooperative jurisdiction. These are hard problems beyond the scope of this paper, but acknowledging them is important. Having examined the gaps, we now turn to how we might fill them – both through technical measures and policy innovations that go hand-in-hand.

5

Defensive Measures and International Collaboration

Mitigating the threats of open-weight LLMs will require more than just regulation. It also calls for technical defenses and cross-border cooperation. Just as AI empowers attackers, it can also empower defenders – a concept often phrased as using "AI to fight AI." Additionally, because open models transcend national boundaries, international collaboration is crucial in setting norms and sharing intelligence on AI-fueled threats. This section explores these dimensions.

5.1. Harnessing AI for Cyber Defense

Advances in AI can be turned toward strengthening cybersecurity. Research and industry efforts are already underway to develop AI-driven defensive tools. Some promising avenues include:

- *AI-Powered Threat Detection:* Machine learning models can analyze network traffic patterns, system logs, or user behavior to detect anomalies that might indicate a breach. Unlike static rules, AI detectors can potentially spot novel attack techniques by learning what "normal" looks like and flagging deviations. For example, an ML model can be trained to detect subtle beaconing from malware or unusual sequences of system calls. Modern anomaly detection systems are incorporating advanced models (including LLMs for log analysis) to catch sophisticated intrusions that evaded traditional signature-based detection [4].
- *Automated Incident Response:* AI can assist in triaging and responding to security incidents. Experimental systems use AI reasoning to interpret security alerts, determine the likely scope of an attack, and even execute containment actions like isolating compromised endpoints or shutting off specific network access – all at machine speed. By automating parts of incident response, organizations hope to react to breaches in seconds or minutes rather than hours, limiting damage. These AI systems can also provide guidance to human analysts, summarizing an ongoing incident and suggesting remediation steps [4].
- *Predictive Threat Intelligence:* Another defensive application is using AI to analyze large volumes of threat data (e.g., threat feeds, forums, past incident reports) to predict emerging attack trends. For instance, AI might identify that threat actors are increasingly discussing a certain software exploit or tactic, allowing defenders to proactively harden that area. Large language models could help synthesize disparate intelligence reports and highlight patterns. The goal is to anticipate attackers' moves – essentially an AI early warning system for new Tactics, Techniques, and Procedures (TTPs) [4].
- *Vulnerability Management:* AI may also help find and fix vulnerabilities before attackers do. By using machine learning to assist code review (like finding security bugs in software code) or by automatically scanning configurations for misconfigurations, AI can make the tedious process of vulnerability assessment more scalable. Some companies are developing AI that suggests patches or reconfigurations as well. If open models can be used by attackers to find vulner-

abilities, defenders should likewise employ AI to discover and remediate them preemptively [4].

However, building these defensive AI systems comes with challenges. Attackers can employ adversarial evasion techniques – essentially designing malware or attack patterns specifically to fool AI detectors (for example, slightly perturbing malicious code or traffic to evade an ML classifier) [4]. The defender's AI also requires vast amounts of high-quality training data (attack data, normal behavior data) which can be hard to obtain and share due to privacy/security concerns [4]. Additionally, AI-based detectors can produce false positives (benign behavior flagged as suspicious), which if too frequent, overwhelm human responders or cause alert fatigue [4]. Despite these issues, the consensus is that we must invest in AI defenses to keep pace with AI-enhanced threats. Policymakers can help by incentivizing R&D in defensive AI. For example, governments could fund the creation of open datasets for cybersecurity (to train AI models), sponsor competitions to develop better AI detection algorithms, or require critical infrastructure organizations to incorporate AI-based monitoring.

Indeed, this paper argues that just as regulators scrutinize offensive AI capabilities, they should also support "AI-enabled defense". The EU AI Act itself, while mainly about risk, could indirectly foster defensive innovation by not placing undue restrictions on it (e.g., clarifying that defensive security AI tools are beneficial and encouraged). More concretely, partnerships between the public and private sector could be formed to share information that helps train defense-oriented models. Greater cooperation is needed among industry peers as well – often companies are hesitant to share attack data due to reputational concerns, but sharing (perhaps via anonymized or aggregated forms) would bolster collective security.

5.2. International Cooperation and Norms

Because open-weight AI and cyber threats ignore national borders, no single nation's policy will suffice. International coordination is key in two areas: (a) cyber threat intelligence sharing and (b) setting common standards and norms for AI security.

Cyber Threat Intelligence (CTI) Sharing: We need robust channels for rapidly sharing information on AI-driven attacks and emerging threats across countries and organizations [4]. This could build on existing cybersecurity alliances. For instance, NATO's Cooperative Cyber Defence Centre of Excellence (CCDCOE) and agencies like the EU's ENISA or the US CISA could spearhead joint efforts to track AI- facilitated threats [4]. If one company or country detects that a certain open-source model is being weaponized in the wild (say, signs of a botnet powered by an LLM generating phishing content), they should promptly alert others. Similarly, new defensive techniques or detection methods for AI attacks should be shared. Establishing an international CTI platform focused on AI threats – akin to how some organizations share indicators of compromise for malware – would help defenders globally stay updated. The timely flow of such intelligence is crucial; delays allow adversaries to reuse tactics elsewhere.

Common Standards for AI Security Evaluation: International collaboration can also produce standards and benchmarks to assess AI models for security risks [4]. The MITRE OCCULT framework we discussed could serve as a foundation – if widely adopted, it provides a consistent way to report how capable a model is in offensive tasks. Perhaps an international body (like ISO or a consortium of AI researchers) could formalize something like an "Offensive AI Capability Index" based on standardized tests. This would let policymakers and companies objectively gauge the risk level of a given model. Regulators could then tie certain measures to the results (for example, if a model crosses a threshold of offensive capability, it might trigger a higher level of scrutiny or safety requirements). Having common evaluation criteria avoids fragmentation – otherwise each country might come up with its own tests, burdening developers and leading to inconsistent risk judgments. Moreover, shared standards mean that if a model is flagged as dangerous in one jurisdiction, others will recognize why. In the same vein, standards for AI robustness (like how resilient a model is to adversarial inputs) and AI transparency (documentation standards like model cards) should be harmonized internationally so that open model publishers worldwide know what is expected.

Joint Research and Development: Nations and institutions can pool resources for research into securing AI. This includes research into techniques like advanced watermarking of AI outputs (so that even if a model is open, maybe its outputs carry a hidden signature that's hard to strip) or

model provenance tracking (methods to verify if a model has been tampered with) [4]. For example, an international grant could fund development of watermarking that is robust against retraining – a challenging task but one that would help if achieved. Another concept is semi-open distribution models [4]: finding a middle ground where models are partly open but with certain sensitive layers or capabilities locked/encrypted, or require a key to unlock (similar to how some dual-license software works). Research could explore if it's possible to split a model's weights such that the base is open but the components that enable the most dangerous functions are separate and only given to verified users. These are innovative ideas that no single company might invest in alone, but collaborative R&D could. Establishing joint AI security labs or testbeds where researchers from different countries simulate AI-driven attack/defense scenarios is another possibility.

Norms of Responsible Behavior: On the diplomatic front, efforts could be made to agree on norms against malicious use of AI – for instance, a norm that states will not knowingly develop or support AI systems for indiscriminate cyberattacks, or that they will cooperate in shutting down egregious misuse. Achieving such norms is difficult (cyber norms in general face challenges, as seen with trying to get consensus on not attacking critical infrastructure). Nonetheless, articulating AI-specific norms can at least set expectations and provide a basis for naming and shaming violators. One could envision something like a pledge (maybe through the UN or another forum) where nations agree to treat the proliferation of offensive AI with the same seriousness as, say, proliferation of certain weapons. Enforcement, of course, remains tricky – nations or groups that see strategic advantage in offensive AI may not abide. But broad international agreement can isolate those actors and justify collective countermeasures.

One concrete step is including AI threats in existing cyber diplomacy dialogs. Groups like the UN's Open-Ended Working Group on ICT security could incorporate language about general-purpose AI and its misuse. The tech industry too can play a role: major AI providers globally might form their own pact (akin to the Cybersecurity Tech Accord) committing to certain safeguards.

In summary, no country can tackle this alone. The open-source community is global, cyber attackers operate globally, and so must our response. From sharing threat intel in real-time to aligning regulations to supporting defensive innovations, collaboration multiplies our defensive capacity, just as open AI multiplied offensive capacity.

Before moving to our final set of recommendations, it's worth noting a subtle point: the widespread availability of offensive AI might paradoxically force more openness in defensive intelligence. When powerful offensive tools are out there for anyone, security can't rely on secret knowledge or siloed responses; it demands collective awareness and agility. In other words, open threats require open (or at least shared) countermeasures.

6

Policy and Technical Recommendations

Building on the analysis above, we propose a multi-faceted approach to mitigate cybersecurity risks from open-weight LLMs while preserving their benefits. The overarching principle is balance: restrictions should be targeted and proportionate to risk, and interventions should be designed to minimize interference with open innovation except where truly necessary. Below is a set of recommendations, divided into categories for clarity.

6.1. Targeted Controls on High-Risk Capabilities

Rather than attempting blunt bans or onerous regulation of entire models, policymakers should focus on specific dangerous capabilities that a model might possess or acquire:

- *Capability-Specific Evaluation & Gating:* Implement evaluation regimes (like OCCULT-style tests) for known high-risk capabilities (e.g. ability to generate malware, break encryption, etc.). If a model is found to consistently demonstrate a particular high-risk capability at a dangerous level, that capability could be "gated" in some way [4]. For example, before releasing a model, developers might remove or weaken certain functionality (if feasible) or use fine-tuning to make the model less competent at that task, without affecting other capabilities. This requires nuance: it is technically challenging to precisely remove one capability without collateral effects. However, the principle is to control use cases rather than the technology broadly. Regulatory guidance could encourage developers to perform such capability audits and document any mitigations taken for risky functions [4].
- *Tiered Access Models:* Explore approaches where the base model is open, but add-ons or fine-tunedversionsthatenableparticularlysensitivetasksarerestricted [4]. For instance, a company might open-release a general LLM, but keep a specialized "cyber exploit generation" fine-tune of it only for licensed security researchers. Alternatively, an API key or user verification could be required to unlock certain features of an otherwise local model (this would need technical enforcement like encrypted weights for those features). While difficult to implement perfectly (since a determined actor could try to recreate the fine-tune), tiered release could slow the proliferation of the most dangerous uses. Policymakers could support pilot programs for such tiered models, and clarify that providing open access to a "safe core" model while locking hazardous add-ons would still count positively in regulatory compliance.
- *Refine Compute Thresholds Criteria:* The AI Act and Code of Practice consider using compute or model size thresholds to identify powerful models. We recommend making this approach more adaptive. Regulators should allow exceptions if a model clearly lacks dangerous capability despite high compute, and conversely, include models that might be small but fine-tuned to be highly potent in a niche offensive area [4]. In practice, this means combine the simple metrics (parameters, FLOPS) with capability evaluations. The law could empower the AI Office to update which models are considered "high-risk GPAI" based on testing and real-world

evidence, not just staticthresholds [4]. This flexibility is crucial to avoid both false negatives (missing a threat from a clever small model) and false positives (overregulating a large but harmless model).

- *AI Red Teaming & Continuous Testing:* Institutionalize programs where independent experts regularly test popular open models for emergent dangerous behaviors (similar to how cybersecurity red teams test systems). The findings should feed into both developers' mitigation actions and regulators' assessments. An international "AI Safety Testing Center" could be established to evaluate models (especially open ones) and publish risk reports. Participating in such testing (and acting on results) could be incentivized or even required for major model releases. This acts as an early warning for capabilities that might need gating or monitoring.

6.2. Strengthening Transparency and Responsibility

Transparency is our best substitute for control in open models. We suggest bolstering responsible disclosure and incentivizing good behavior:

- *Mandatory Model Documentation ("Responsible Labeling"):* Require that any release of a powerful open-weight model be accompanied by a detailed Model Card and Risk Assessment document [5]. This documentation should explicitly state what the model can and cannot do (to the best of the developer's knowledge), including any offensive capabilities identified through testing. For example, if OCCULT tests show the model can produce malware code with 80% success, that should be clearly disclosed. Think of it as a "nutrition label" for AI models. While the current GPAI Code of Practice encourages this, making it a norm (or requirement for those seeking certain protections) would improve industry standards [4].
- *Pre-Release Safety Evaluations and Disclosure:* Encourage or mandate that open model developers perform safety tests (red teaming, bias/harm assessment, dual-use assessment) before release and publish the results [5]. If dangerous behaviors are found, describe what mitigations were attempted (if any) or why the release is still justified. This level of transparency serves two purposes: it informs downstream users and provides regulators evidence that due care was taken (important if something goes wrong later). It's analogous to clinical trial data for a new drug – even if it's ultimately made available, the evaluation data must be out there.
- *Partial Weight Release or Mechanisms for Friction:* Research and consider methods where models can be released in a form that imposes some friction on misuse. For example, one idea is "selective weight obfuscation": intentionally obfuscating or encrypting parts of the model weights that are responsible for very specific high-risk knowledge (if identifiable), requiring an extra step to activate [4]. Another idea is rate-limited APIs or compute puzzles integrated into models – though these can likely be removed by an expert, they might deter casual misuse. The point is to discuss innovative ways to make misuse slightly harder without crippling beneficial use. Policymakers could run challenges or grants for such techniques. Any such method should be openly documented (so users know what was done) and ideally open-source, to align with trust principles.
- *Liability Safe Harbors & Positive Incentives:* A stick-only approach may backfire. Therefore, introduce safe harbor provisions or incentives for open model providers who follow best practices [4]. For instance, if a company releases a model openly but complies with a recognized Code of Practice (detailed documentation, testing, etc.), they could be shielded from certain liabilities or enforcement actions, on the condition that any unforeseen harm was truly beyond their control. This is similar to how some jurisdictions handle responsible disclosure – encourage it by offering legal safe harbor. Additionally, governments could provide tax credits, grants, or awards for organizations that significantly contribute to defensive AI or open transparency. If an open model release includes an unprecedented level of transparency and safety measures, that should be applauded and set as a benchmark, not punished. Conversely, those who release powerful models with zero documentation or recklessly could face penalties (or at least reputational consequences) as negligence.
- *Clarify "Open Source" Definition for AI in Regulations:* Regulators should explicitly define what qualifies as an open-source AI model in the context of exemptions [4]. We recommend a pragmatic definition that recognizes the reality of open-weight models (weights open, data possibly

not), and tie the exemption to compliance with documentation and transparency rather than a strict OSI license criterion. In other words, say: if you release a model freely for use, you won't be subject to full high-risk obligations provided you supply the necessary documentation and the model is not foreseeably used in prohibited domains. This removes ambiguity and ensures that open model developers know how to remain on the right side of the law [4, 5]. The EU AI Office, for example, could issue guidance that interprets the Act's open-source clause in this way.

- *Shift Regulatory Focus Downstream:* Regulators and enforcement agencies should prioritize monitoring the use of models in deployed applications, rather than the mere release of models [4]. This means focusing on sectors where AI is applied (like healthcare, finance, etc.) to ensure if someone plugs an open model into a high-risk system, that end system is compliant and safe. If something goes wrong with an AI-driven medical device, blame should lie with the deployer who chose and configured the AI, not automatically the model creator. This approach aligns resources with actual impact and encourages responsible integration. Concretely, it could involve sectoral regulators (e.g., medical device regulators, aviation safety, etc.) adding checks for AI components in certified systems.

6.3. Bridging Open Innovation and Security Oversight

Finally, a few broader actions to bridge the gap between the open AI community and the security policy community:

- *Community Engagement and Self-Governance:* The open-source AI community (researchers, developers on platforms like Hugging Face, etc.) should be engaged in developing security norms. For example, encourage the creation of an Open AI Security Council drawn from leading open model contributors, to issue guidelines or review major releases. Much like how open-source software communities have committees for handling security vulnerabilities (like the Linux Foundation's security groups), the AI community could self-organize to preempt government mandates by showing responsibility. Policymakers could support this by providing fora for dialogue and ensuring that community-developed standards are recognized in regulation as acceptable compliance routes.
- *Education and Best-Practice Sharing:* Many actors releasing or using open models may simply not be fully aware of the risks or best practices. Initiatives to educate open-source developers on cybersecurity (for instance, publishing easy-to-follow guides on how to safely fine-tune a model without stripping safety layers, or how to evaluate your model for dual-use concerns) would raise the bar overall. Government agencies or universities could partner with open-source platforms to disseminate such guidance. Possibly integrate some security checklist into model- sharing hubs (e.g., when someone uploads a model, prompt them with "Did you evaluate X? Here's how.").
- *Monitoring and Accountability for Abuse:* Develop ways to monitor how open models are being used maliciously, without infringing privacy unduly. This could involve setting honeypots (e.g., fake torrent or forum posts advertising "illegal AI models" to observe interest and tactics) or funding threat intelligence focused on AI misuse trends. When clear cases of abuse are found (like a new Xanthorox-type tool), there should be mechanisms to hold those actors accountable – through law enforcement if criminal, or at least by disrupting their infrastructure (the way botnet takedowns are done). The message should be that openly releasing a model for good purposes is supported, but releasing or using models explicitly for cybercrime will be pursued and penalized. This is an analog to how we treat dual-use technology in other fields (e.g., publishing encryption tools is legal, using them for crime is not – but if someone sells a tool specifically to commit crimes, they may face charges).
- *International Agreements for AI in Cyber Warfare:* Push for integration of AI considerations in international cyber warfare norms. For instance, countries could agree that fully autonomous offensive AI (that targets indiscriminately) is off-limits – similar to how discussions are ongoing about lethal autonomous weapons. While hard to verify, getting such commitments can shape military and intelligence community behavior and signal that even nations see uncontrolled offensive AI as dangerous to all. At the very least, it opens channels for dialogue to avoid

unintended escalation (e.g., if an AI-generated attack is detected, nations should talk before assuming it was a deliberate state act – since attribution is even harder with AI in the mix).

Each of these recommendations requires efforts from multiple stakeholders – governments, international bodies, companies, open-source developers, and researchers. It truly is a collective action problem. But the benefit of open models is that collective action is possible: many minds globally can contribute to solutions, just as many can inadvertently contribute to the problem.

By implementing these measures, we can aim to create an ecosystem where open-weight LLMs are developed and used in a responsible manner: where innovation continues in the open, but with guardrails and accountability, and where defenders are as empowered as attackers.

7

Conclusions

The rise of open-weight large language models marks a pivotal moment in both AI development and cybersecurity. On one hand, open models like LLaMA, Mistral, and DeepSeek-R1 have democratized access to cutting-edge AI capabilities, spurring remarkable innovation and community collaboration. On the other hand, as this article has detailed, the unfettered availability of such models has introduced new avenues for cyber offense – effectively open-sourcing the means of cyberattacks. Advanced skills that were once the domain of elite hackers or state agencies can now be emulated or enhanced by AI, potentially enabling less skilled actors to punch above their weight. This paradigm shift in the threat landscape calls for an equally adaptive shift in our defensive and policy approaches.

Current regulatory frameworks, exemplified by the EU AI Act, provide a starting point but are not fully equipped for the unique challenges of open-weight AI. The AI Act's risk-based approach and high-level principles are sound, yet the specifics – such as the open-source exemption and obligations tied to control – require careful interpretation to avoid both regulatory overreach and loopholes. As we've argued, a key adaptation is to formalize the transfer of responsibilities: when an open model is misused or modified, the accountability should follow the actor who wields or alters the model, not invariably fall on the original provider. This nuanced view is essential to maintain fairness and efficacy in enforcement. The consultation insights we integrated underscore this point, urging clarity on concepts like substantial modification (e.g., treating the removal of safety features as a trigger for new obligations) and emphasizing the "mitigation gap" that appears once model weights are public [3, 5].

Policymakers should recognize that traditional mitigation strategies must be re-imagined for open AI. Security can no longer rely on keeping models behind closed APIs or closely monitoring every output – strategies that work for proprietary systems but fail when anyone can run the AI offline. Instead, the focus must shift to preemptive and distributed measures: ensuring models are released with safeguards and knowledge of their risks, and empowering the ecosystem at large (downstream developers, users, and defenders) to act responsibly. In practical terms, this means beefing up transparency requirements and fostering a culture of "responsible open release". It also means regulators might need to get more technical – engaging with AI experts to set testing standards and not shy away from adjusting rules as the technology evolves. The AI Act's future implementation via the AI Office and codes of practice will be crucial; done right, it can serve as a blueprint globally for governing general-purpose AI without hampering open science.

From a technical and strategic perspective, mitigating cyber risks in open-weight LLMs is not solely a regulatory issue, but a socio-technical one. We must invest in defensive technology – using AI to safeguard systems – with the same vigor that adversaries invest in offensive AI. The battle is not AI vs. human, but AI-empowered defense vs. AI-empowered offense. This calls for a recalibration of resources and priorities in cybersecurity. Organizations and governments should incorporate AI threat scenarios in their planning and exercises. Cybersecurity professionals may need training on AI, and vice versa, AI developers need exposure to security thinking. Bridging these communities is imperative; interdisciplinary collaboration (e.g., between AI researchers and cyber threat analysts) can yield novel solutions neither would devise alone.

Internationally, a collective stance on AI in cybersecurity can set norms and reduce the chances of an AI- fueled “arms race” in cybercrime. While criminals won’t sign accords, states and companies can commit to certain baselines – for instance, not releasing models without due diligence, and cooperating to counter malicious uses. If one major jurisdiction enacts sensible measures, others should harmonize to avoid regulatory arbitrage where bad actors move to laxer environments. The global nature of open- source means we are only as strong as the weakest link in terms of regulatory safe havens for abuse. Thus, ongoing dialogue from the G7 to the UN, and in multi-stakeholder forums, should integrate these AI-cyber risk topics.

In closing, open-weight LLMs present a classic dual-use dilemma: the same technology that can benefit humanity can also cause harm. We have faced similar challenges before – from cryptography to biotechnology – and learned that outright suppression is neither practical nor desirable. The way forward lies in smart risk management: identify the truly high-risk aspects and intervene there, support the development of countermeasures, and cultivate a norm of responsibility among practitioners. Open collaboration need not be the enemy of security; in fact, open communities can often self-police and innovate solutions faster than closed ones, given the right guidance and incentives.

The recommendations outlined in this paper aim to chart a course where we neither naïvely ignore the threats nor needlessly hamper progress. It is a path of pragmatic equilibrium – acknowledging that open models are here to stay and can be forces for good, but also insisting that certain guardrails and accountability mechanisms evolve alongside them. This balancing act is undoubtedly challenging, but it is one we must strive to achieve. The security of our digital society may depend on our ability to adapt policy and practice to this new era of AI ubiquity.

Future Outlook: As a final thought, it’s worth noting that technology will continue to advance – tomorrow’s open AI models might include multimodal systems (combining vision, text, and more) or even autonomous agent frameworks far more capable than today’s. Preparing for those now, by establishing robust principles and frameworks, will pay off down the line. Continued research is needed in areas like AI model watermarking, federated model usage tracking, adversarial robustness, and more, to technically underpin policy aims. Moreover, continued observation of the impacts of current open models will inform whether our mitigation steps are working or if we need course corrections. This is an iterative process. The hope is that through scholarly analysis, industry best practices, and enlightened policymaking, we can steer the age of open-weight LLMs towards a future where security and openness co-exist – where the benefits of these powerful tools can be enjoyed widely, without ushering in a wave of uncontrolled cyber mayhem.

References

[1] California Senate. “SB-1047 Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.” In: (2024). URL: https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047.

[2] European Commission. “Third Draft of the General-Purpose AI Code of Practice: Commitments by Providers of General-Purpose AI Models with Systemic Risk - Safety and Security Section. Draft Document.” In: (2025). URL: https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts.

[3] European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).* OJ L, 2024/1689, June 2024. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

[4] Alfonso de Gregorio. *Mitigating Cyber Risk in the Age of Open-Weight LLMs: Policy Gaps and Technical Realities*. 2025. arXiv: 2505.17109 [cs.CR]. URL: https://arxiv.org/abs/2505.17109.

[5] Alfonso De Gregorio. *Consultation input from Alfonso De Gregorio to the European Commission on the AI Act (2025). Specific clarifications on Articles 9-15 and substantial modification (unpublished consultation response content)*. 2025.

[6] M. Kouremetis et al. “OCCULT: Evaluating Large Language Models for Offensive Cyber Operation Capabilities.” In: (Apr. 2025). arXiv: cs.CR/arXiv:2502.15797. URL: https://doi.org/10.48550/arXiv.2502.15797.

[7] E. Montalbano. “Autonomous, GenAI-Driven Attacker Platform Enters the Chat”. In: *Dark Reading* 12 (2025), p. 2025. URL: https://www.darkreading.com/threat-intelligence/autonomous-genai-attacker-platform-chat.

[8] OSI. *The Open Source AI Definition – 1.0*. Apr. 2024. URL: https://opensource.org/ai/open-source-ai-definition.

[9] M. Rodriguez et al. “A Framework for Evaluating Emerging Cyberattack Capabilities of AI”. In: (Mar. 2025). URL: https://doi.org/10.48550/arXiv.2503.11917.

[10] *Xanthorox AI – The Next Generation of Malicious AI Threats Emerges*. 2025. URL: https://slashnext.com/blog/xanthorox-ai-the-next-generation-of-malicious-ai-threats-emerges/.