

# Access Controls Will Solve the Dual-Use Dilemma

Evžen Wybitul<sup>1</sup>

## Abstract

AI safety systems face a dual-use dilemma. The same request can be either harmless or harmful depending on who made it and why. Thus, if the system makes decisions based solely on the request’s content, it will refuse some legitimate queries and let harmful ones pass. To address this, we propose a conceptual access control framework, based on verified user credentials (such as institutional affiliation) and classifiers that assign model outputs to risk categories (such as advanced virology). The system permits responses only when the user’s verified credentials match the category’s requirements. For implementation of the model output classifiers, we introduce a theoretical approach utilizing small, gated expert modules integrated into the generator model, trained with gradient routing, that enable efficient risk detection without the capability gap problems of external monitors. While open questions remain about the verification mechanisms, risk categories, and the technical implementation, our framework makes the first step toward enabling granular governance of AI capabilities: verified users gain access to specialized knowledge without arbitrary restrictions, while adversaries are blocked from it. This contextual approach reconciles model utility with robust safety, addressing the dual-use dilemma.

## 1. Introduction

User requests — and with them, model outputs — exist on a spectrum from clearly benign to clearly harmful, with most falling in the grey zone in the middle (example in Figure 1). In the grey zone, the same output could be considered harmful or harmless, depending not on its content, but on its *real-world context*: who requested it and for what purpose.

Safety systems that rely solely on content analysis immediately face the *dual-use dilemma*. Since the same request

<sup>1</sup>ETH Zurich, Switzerland. Correspondence to: Evžen Wybitul <wybitul.evzen@gmail.com>.

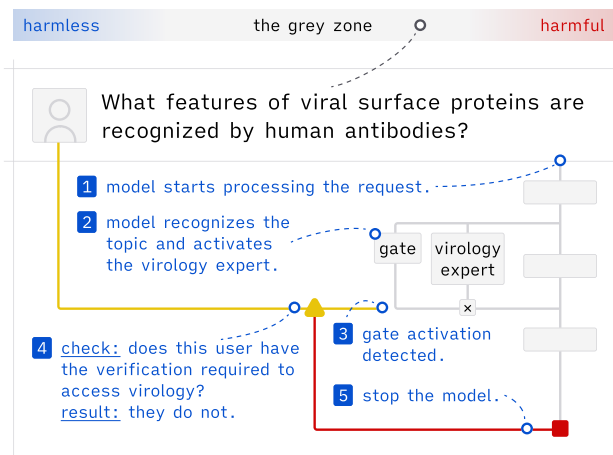


Figure 1. The user is asking a question from the grey zone: a question that could be either harmless or harmful, depending on its real-world context. The schema shows how the system we propose would handle it. (1) The model is trained to be helpful and begins to answer the question. (2) During the forward pass, the model activates its virology expert module because it is relevant to the question. (3) The activation of the expert is observed by an external mechanism that immediately (4) checks in the company’s database if the user has the required authorization to access virology knowledge. (5) Since they don’t, the model is stopped. If they did, the model would be allowed to give an answer.

can be either harmless or harmful depending on the context, wherever they draw the refusal line, they will restrict model utility for legitimate users while letting slip harmful requests from adversaries. Some safety systems try to address this by considering real-world context alongside content. However, they typically infer the context from the content itself, making it easy for adversaries to fabricate.

In this paper, we argue that informative, hard-to-fabricate real-world context could be obtained using user-level verifications such as institutional affiliation, or know-your-customer checks. We then address the dual-use dilemma with two contributions:

1. We show how this type of context could be used jointly with content analysis in a safety system based on access controls (Lampson, 1974). First, generated content

---

would be classified into risk categories. Then, a check would be performed to see whether the user has the verifications required to access the detected categories.

2. We propose a novel technical approach to risk category classification that is based on gradient routing (Cloud et al., 2024). Our proposal avoids having the capability gap between a model and its monitors that can make output monitoring methods non-robust (Jin et al., 2024).

Our framework is a first step toward solving the challenge of “detection and authorization of dual-use capability at inference time” that was highlighted by a recent survey of problems in technical AI governance (Reuel et al., 2025) and also raised by the U.S. AI Safety Institute (2024). As such, it has important governance implications, potentially enabling a more nuanced regulatory approach where access to powerful AI capabilities is stratified rather than binary, with policies that differentiate between user types and user contexts rather than focusing solely on model capabilities. The choice of appropriate verification mechanisms and risk categories remains for future work and should ideally happen jointly with stakeholders from academia, AI governance, and industry. Nevertheless, our approach offers a promising direction for addressing the dual-use dilemma.

## 2. Current Safety Methods Don’t Solve the Dual-Use Dilemma

We evaluate three approaches from the AI safety literature to see how sensitive they are to contextual information, and whether their sources of real-world context are trustworthy — that is, hard to manipulate by an adversary.

First, to illustrate the need for context, consider decomposition attacks (Glukhov et al., 2023; 2024): transforming a clearly harmful query, such as “How to modify a virus to avoid immune detection?”, into a series of mundane technical questions, like the “What features of viral surface proteins are recognized by human antibodies?” from Figure 1. Here, the attacker exploits the dual-use dilemma, and the fact that model providers cannot refuse grey zone requests to preserve model utility.

### 2.1. Unlearning: Non-Contextual Removal of Concepts

Unlearning methods aim to remove specific knowledge, concepts, or capabilities from a model after training (Liu et al., 2024). Their goal is to eliminate the model’s ability to generate harmful content while preserving other capabilities.

Unlearning faces significant technical challenges even for preventing behaviours that are clearly harmful. As noted by Cooper et al. (2024) and Barez et al. (2025), capabilities are hard to define, hard to remove without side effects, and

it is hard to trace them back to specific data points. Many unlearning approaches mask rather than truly remove the targeted knowledge (Deeb & Roger, 2025). Moreover, even nascent robust unlearning methods (Cloud et al., 2024; Lee et al., 2025) are not contextual, and thus don’t address the dual-use dilemma without additional assumptions.

### 2.2. Safety Training: The Model Reacts to Context

Safety training methods modify the model’s training process to align its outputs with human preferences. This category includes safety pre-training (Maini et al., 2025), RLHF (Christiano et al., 2023), and safety finetuning.

Unlike unlearning, these methods are contextual. They don’t remove capabilities entirely but train the model to selectively deploy them based on, among other things, the perceived legitimacy and harmlessness of the request. However, these qualities are entirely inferred from content supplied by the user, such as the request content, the chat history, or the model’s memories about past conversations. It should be no surprise, then, that models are susceptible to attacks that fabricate in-chat context (Zeng et al., 2024), or attacks that diminish models’ sensitivity to in-chat context, e.g. through multi-round escalation (Russinovich et al., 2025). Without access to trustworthy real-world context of the request, the model cannot make truly informed decisions about grey zone requests, and thus cannot robustly address the dual-use dilemma.

### 2.3. Post-Processing: External Systems React to Context

Post-processing methods are systems that classify user inputs and model outputs for the purposes of steering the underlying model, and monitoring and filtering its outputs. Sometimes, these methods are used for usage monitoring, as is the case with Anthropic’s Clio (Tamkin et al., 2024; Handa et al., 2025), other times, they are used for safety, as with Llama Guard (Inan et al., 2023) and Constitutional Classifiers (Sharma et al., 2025). However, similarly to safety training, the “real-world” context these methods work with is currently inferred mostly from user-supplied content and is thus untrustworthy and vulnerable to attacks, as evidenced by the many jailbreaks that successfully target current production systems (Zhang et al., 2025). Nevertheless, these methods could be modified to incorporate external contextual information, potentially serving as a foundation for more trustworthy, contextual safety mechanisms. We discuss this option in Section 3.4.2.

## 3. Access Controls as a Solution

Current safety systems face the dual-use dilemma because they lack trustworthy information about who is making a request and why. In this section, we describe an access

---

control system that addresses this problem by verifying user credentials before granting access to sensitive knowledge.

### 3.1. Overview of the Access Control Framework

We propose a defensive system where grey-zone requests are refused by default, but users can gain access to specific categories of knowledge if they undergo verification.

When a model provider sets up the system, they make two design decisions with the help of domain experts. First, they define **content categories** (Section 3.2): groups of sensitive topics organized by domain and risk rating. Second, for each content category, they specify a **verification mechanism** (Section 3.3): the verification process users must complete to access that category.

Whenever the model generates an output, the system performs **content classification** (Section 3.4) to check if the model’s output belongs to any pre-defined content category. If the user lacks authorization for the detected category, the system implements appropriate **system responses** (Section 3.5) ranging from enhanced logging to full refusal.

For example, in biology, basic knowledge and common techniques would remain freely accessible, widespread techniques like CRISPR might require ID-based verification, and dangerous techniques like aerosolization might require government biosafety certifications. If a user asks for help with CRISPR laboratory protocols, the system would detect that the request belongs to the low-risk category, check whether the user has verified their ID, and either provide the information or ask them to complete verification first.

This approach directly addresses both sides of the dual-use dilemma. Decomposition attacks become harder because the system refuses grey-zone requests by default—attackers would need legitimate credentials rather than clever prompting. Simultaneously, verified users can access specialized knowledge that would face blanket restrictions under current approaches.

The main concern is increased user friction, but we show in Section 4 that this should be minimal; among other things, because most users will never make grey-zone requests.

### 3.2. Content Categories

Content categories are groups of sensitive topics organized by domain and risk level, which the model providers will develop with domain experts on top of existing risk frameworks. For example, in biology, they could use the biosafety levels (BSL) (Centers for Disease Control and Prevention & National Institutes of Health, 2020) and dual-use research of concern policies (United States Government, 2012).

However, these existing frameworks only categorize entities like organisms or compounds that are too broad for content

control in our system. Thus, experts would need to decompose these high-level concepts into component knowledge pieces. For instance, for growing and spreading a dangerous BSL-3 pathogen, this might include specific procurement methods, cultivation techniques, purification methods, and protocols for specialized equipment. For each component, experts would assess how many harmless versus harmful applications it enables, then assign it to an appropriate risk category: widely useful techniques with many benign applications, such as using CRISPR, would remain in low-risk categories with minimal verification, while specialized techniques with limited legitimate uses, such as aerosolization of bacteria, would require higher verification levels.

Evidence from chemistry suggests this approach could work: the risk schedules of the Chemical Weapons Convention already identify not just controlled compounds but their precursors or specific equipment (Organisation for the Prohibition of Chemical Weapons, 1993), which is a form of concept decomposition. Nevertheless, some harmful applications might not decompose so neatly; we discuss this limitation in Section 4.

### 3.3. Verification Mechanisms and Credentialling Systems

Verification mechanisms are the credential processes that users must complete to access each content category. The system will initially vary across model providers, but we expect it to follow a three-tier structure: most content will require no verification, content categories with moderate risk will require basic identity checks or institutional affiliation, and high-risk categories will require domain-specific certifications. In each case, model providers will build on existing verification infrastructure rather than creating new systems, consulting domain experts to identify appropriate mechanisms for each field.

For low-risk content categories, model providers could use established identity verification services like Stripe Identity (Stripe, Inc., 2024) or institutional systems like ORCID (ORCID, Inc., 2024) which provide global, standardized, low-friction solutions with one-time costs under \$2 per user. These systems would serve primarily to maintain audit trails for post-incident investigation and provide a deterrent effect, rather than as security barriers for truly dangerous knowledge.

High-risk content categories could leverage existing domain-specific certifications that indicate users’ ability to handle sensitive information and materials responsibly. Model providers would work with domain experts and national authorities to identify appropriate certifications, adapting existing physical-world credential systems to knowledge access control. For biological content categories, for example, the system might draw on governmental certifications for

handling high-BSL organisms we mentioned in Section 3.2, and analogous certifications in countries outside the US.

Governance of verification systems, including the requirements and the appeals processes, will initially vary across providers. Over time, successful approaches may inform industry coordination and eventual standardization, similar to how content moderation and know-your-customer standards evolved.

This approach faces several limitations. For high-risk categories, relying on existing certifications may be overly restrictive, potentially excluding some users who should have access. However, we argue in Section 4 that knowledge in high-risk categories would likely face blanket restrictions anyway, and our approach enables access for at least some verified users instead of none. In the same section, we discuss open problems including equity concerns for users in developing countries and privacy implications.

### 3.4. Implementing Content Classification

A core requirement of the verification-based access control system described in Section 3 is being able to reliably classify model outputs into risk categories. This classification needs to address key challenges: accuracy with minimal false positives, resistance to adversarial attacks, and efficiency. We examine two approaches to implementing this classification — one currently available and one theoretical — and discuss their trade-offs.

#### 3.4.1. POST-PROCESSING

As discussed in Section 2, current systems already rely on post-processing classifiers that analyse outputs before delivery to users. These could be adapted for content classification into risk categories in an access control system. For example, a classifier could be trained to identify moderately advanced virology topics and, if detected, could trigger verification of user permissions before delivering the output.

The key advantage of post-processing systems is modularity, as they can be developed and updated independently of the generation models they oversee. However, they face a trade-off between usability (latency) and safety (Kumar et al., 2025): prioritizing low latency can create a capability gap between generators and monitors that sophisticated language models can exploit (Jin et al., 2024). Despite this limitation, recent post-processing methods show acceptable efficiency and resilience toward jailbreaks (Sharma et al., 2025) and could provide a practical initial implementation path for output-based access control.

#### 3.4.2. GRADIENT ROUTING

To address the capability gap problem in post-processing methods, we detail how gradient routing (Cloud et al., 2024)

could be adapted to classify model outputs into risk categories. Our adaptation represents a theoretical direction for integrating risk category detection directly into model architecture. This approach can be combined with post-processing methods and offers different trade-offs.

We propose augmenting models with small expert modules controlled by learned gates, as shown in Figure 1. During model training, assuming we have examples from the relevant risk categories (e.g., advanced virology), the expert module would be made to receive gradients exclusively from these examples. Simultaneously, the module’s gate would be trained to activate the module on these examples using an auxiliary loss function. We hypothesize this would lead to the module obtaining specialist knowledge from the risk domain, and to the model learning to activate the module whenever this knowledge is needed.

Figure 1 illustrates this in the context of the wider access control system: when a user requests information about viral surface proteins, the model correctly recognizes the domain and activates the virology gate, engaging the expert but also triggering the verification system. If the user lacks appropriate credentials, the system takes the appropriate response — here, halting the generation.

This approach eliminates the efficiency-capability trade-off inherent in post-processing methods: the risk category classifier is embedded within the larger model, so there is no capability gap, and because the gated module is small and integrated into the model’s forward pass, it is computationally efficient with minimal impact on latency.

Crucially, during standard training, the model is naturally incentivized to identify risk category information in inputs and surface this in its representation space, as this is directly linked to enabling the right specialized module at the right time. This approach differs fundamentally from post hoc methods like probing or sparse autoencoders (Cunningham et al., 2023), which cannot guarantee the model actually operates with the concepts they attempt to identify.

Additionally, it resolves the adversarial dynamic (model vs. monitors, helpful vs. harmless) common in AI safety by aligning training incentives with safety objectives.

Several technical challenges remain. While gradient routing has shown promise in creating specialized modules in smaller models, its effectiveness in larger language models, particularly with the gated-expert architecture, requires empirical validation. Other key challenges include preventing false positives and false negatives, though regularization techniques and adjusted detection thresholds, respectively, could mitigate these issues.

Our approach also requires identifying risk categories during initial training, prompting research into adaptation of



---

gradient routing for fine-tuning scenarios. Despite these challenges, the approach offers promising theoretical properties that warrant experimental investigation.

### 3.5. System Responses

System responses — the way the system reacts when a user makes a request they do not have the authorization to make — can vary based on risk category and the confidence of the classification. In a way, this can balance out the friction added by difficult-to-classify topics or imperfect classifiers.

The initial implementation might use a three-phase response system: (1) for outputs belonging to a risk category with high-confidence, immediate refusal with a specific explanation of the verification required; (2) for borderline classifications, continued generation, but with enhanced logging and additional safety review using more costly post-processing methods.

## 4. Feasibility and Limitations

### User friction analysis.

- Access controls introduce friction from two sources: intentional verification requirements for grey-zone requests and accidental false positives from imperfect classification.
- By design, all grey-zone requests require verification even with perfect classifiers. Additionally, classification errors may incorrectly flag clearly harmless requests (e.g., high-school biology as advanced virology).
- To understand the combined impact, consider that under 3% of requests involve biology topics according to the Anthropic Economic Index (Handa et al., 2025) (see Appendix A for more details).
- Even if ALL biology requests required verification—an extreme upper bound—this would affect fewer users than current system friction. Existing safety systems refuse 1–7% of benign requests as false positives, with Claude achieving the best rate of 0.4%. Requiring verification for all biology (2.43%) would approximately double Claude’s friction rate.
- Additionally, verification friction differs qualitatively from current false positives—users can resolve issues through one-time credential verification rather than facing permanent refusal.
- Companies can calibrate friction through multiple mechanisms: adjusting grey-zone boundaries, tuning classifier thresholds, and implementing graduated responses (requesting clarification, additional context, or

secondary review) rather than immediate verification requirements.

- Companies can determine optimal settings empirically through: (1) internal red-team decomposition attacks on concerning capabilities, (2) gradual rollout with logging to measure user impact, and (3) iterative adjustment based on safety-friction tradeoffs.

### Developer incentives for adoption.

- Access controls enable competitive advantages by allowing companies to serve “dark-grey” requests that competitors refuse for safety reasons, while adding surgical restrictions to “light-grey” requests vulnerable to decomposition attacks.
- Without access controls, continued decomposition attack success will likely trigger broad government regulations restricting model capabilities entirely. Surgical access controls allow compliance with safety requirements while preserving advanced capabilities for verified users, creating market differentiation.
- Even without perfect industry coordination, first movers gain competitive advantages in serving previously restricted capabilities.

This approach faces two key limitations. First, developing countries may lack advanced certification infrastructure, potentially preventing access to knowledge. This is a challenge that requires international cooperation.

Also, privacy concerns.

Also, what if concepts don’t decompose neatly and don’t have a nice harmful bottleneck?

Existing certifications may be overly broad for our needs. For example, biosafety certifications verify physical equipment for handling pathogens, which is overly strict for our simpler use case of restricting access to knowledge. We concede that the existing certification systems are just an imperfect proxy, but we believe they would be iteratively refined as the technology matures. We also note that current legislation like the EU AI Act [cite] favours broad capability restrictions for high-risk domains, which means that our system, however imperfect, might enable access for at least some users where there would be none under the status quo.

## 5. Conclusion

We argued that safety systems that do not utilize contextual information face a lose-lose *dual-use dilemma*: they will restrict model utility for some legitimate users while still allowing some adversaries to use the model for ill. To

address this problem, we introduced a new access control framework that limits access to outputs from certain risk categories only to users with relevant verifications (which serve as proxies for trustworthy real-world context). We also proposed a novel technical solution for classifying outputs into risk categories based on gradient routing that has the potential to resolve the efficiency-robustness trade-off of post-processing methods.

Beyond addressing specific technical challenges, our framework represents a promising governance shift from working with model-level abstractions and binary capability restrictions toward more granular user-level access controls. This offers a practical pathway for regulating increasingly powerful AI systems through stratified access rather than blanket capability limitations.

## Acknowledgements

We thank Jakub Kryś, and Dennis Akar for their feedback on a draft of this paper. We thank Joseph Miller, Alex Cloud, Alex Turner, and Jacob Goldman-Wetzler for discussions on gradient routing.

## References

- Barez, F., Fu, T., Prabhu, A., Casper, S., Sanyal, A., Bibi, A., O’Gara, A., Kirk, R., Bucknall, B., Fist, T., Ong, L., Torr, P., Lam, K.-Y., Trager, R., Krueger, D., Mindermann, S., Hernandez-Orallo, J., Geva, M., and Gal, Y. Open problems in machine unlearning for ai safety, 2025. URL <https://arxiv.org/abs/2501.04952>.
- Centers for Disease Control and Prevention and National Institutes of Health. Biosafety in microbiological and biomedical laboratories. Technical report, U.S. Department of Health and Human Services, Atlanta, GA, 2020. URL <https://www.cdc.gov/labs/BMBL.html>. Defines Biosafety Levels (BSL-1 through BSL-4) used in the United States.
- Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, 2023. URL <https://arxiv.org/abs/1706.03741>.
- Cloud, A., Goldman-Wetzler, J., Wybitul, E., Miller, J., and Turner, A. M. Gradient routing: Masking gradients to localize computation in neural networks, 2024. URL <https://arxiv.org/abs/2410.04332>.
- Cooper, A. F., Choquette-Choo, C. A., Bogen, M., Jagielski, M., Filippova, K., Liu, K. Z., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., Shumailov, I., Triantafillou, E., Kairouz, P., Mitchell, N., Liang, P., Ho, D. E., Choi, Y., Koyejo, S., Delgado, F., Grimmermann, J., Shmatikov, V., Sa, C. D., Barocas, S., Cyphert, A., Lemley, M., danah boyd, Vaughan, J. W., Brundage, M., Bau, D., Neel, S., Jacobs, A. Z., Terzis, A., Wallach, H., Papernot, N., and Lee, K. Machine unlearning doesn’t do what you think: Lessons for generative ai policy, research, and practice, 2024. URL <https://arxiv.org/abs/2412.06966>.
- Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, 2023. URL <https://arxiv.org/abs/2309.08600>.
- Deeb, A. and Roger, F. Do unlearning methods remove information from language model weights?, 2025. URL <https://arxiv.org/abs/2410.08827>.
- Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., and Pappan, V. Llm censorship: A machine learning challenge or a computer security problem?, 2023. URL <https://arxiv.org/abs/2307.10719>.
- Glukhov, D., Han, Z., Shumailov, I., Pappan, V., and Papernot, N. Breach by a thousand leaks: Unsafe information leakage in ‘safe’ ai responses, 2024. URL <https://arxiv.org/abs/2407.02551>.
- Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., and Ganguli, D. Which economic tasks are performed with ai? evidence from millions of claude conversations, 2025. URL <https://arxiv.org/abs/2503.04761>.
- Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL <https://arxiv.org/abs/2312.06674>.
- Jin, H., Zhou, A., Menke, J. D., and Wang, H. Jailbreaking large language models against moderation guardrails via cipher characters, 2024. URL <https://arxiv.org/abs/2405.20413>.
- Kumar, D., Birur, N. A., Baswa, T., Agarwal, S., and Harshangi, P. No free lunch with guardrails, 2025. URL <https://arxiv.org/abs/2504.00441>.
- Lampson, B. Protection. *ACM SIGOPS Operating Systems Review*, 8:18–24, 01 1974. doi: 10.1145/775265.775268.
- Lee, B. W., Foote, A., Infanger, A., Shor, L., Kamath, H., Goldman-Wetzler, J., Woodworth, B., Cloud, A., and Turner, A. M. Distillation robustifies unlearning, 2025. URL <https://arxiv.org/abs/2506.06278>.

- Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C. Y., Xu, X., Li, H., Varshney, K. R., Bansal, M., Koyejo, S., and Liu, Y. Rethinking machine unlearning for large language models, 2024. URL <https://arxiv.org/abs/2402.08787>.
- Maini, P., Goyal, S., Sam, D., Robey, A., Savani, Y., Jiang, Y., Zou, A., Lipton, Z. C., and Kolter, J. Z. Safety pre-training: Toward the next generation of safe ai, 2025. URL <https://arxiv.org/abs/2504.16980>.
- ORCID, Inc. ORCID: Connecting research and researchers, 2024. URL <https://orcid.org/>. Global, persistent identifier system for researchers and scholars.
- Organisation for the Prohibition of Chemical Weapons. Convention on the prohibition of the development, production, stockpiling and use of chemical weapons and on their destruction, 1993. URL <https://www.opcw.org/chemical-weapons-convention>. Contains the Annex on Chemicals with Schedules 1–3.
- Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Bau, D., Bricman, P., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. Open problems in technical ai governance, 2025. URL <https://arxiv.org/abs/2407.14981>.
- Russinovich, M., Salem, A., and Eldan, R. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2025. URL <https://arxiv.org/abs/2404.01833>.
- Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., Ong, E., Peng, A., Agarwal, R., Anil, C., Asbell, A., Bailey, N., Benton, J., Bluemke, E., Bowman, S. R., Christiansen, E., Cunningham, H., Dau, A., Gopal, A., Gilson, R., Graham, L., Howard, L., Kalra, N., Lee, T., Lin, K., Lofgren, P., Mosconi, F., O’Hara, C., Olsson, C., Petrini, L., Rajani, S., Saxena, N., Silverstein, A., Singh, T., Summers, T., Tang, L., Troy, K. K., Weisser, C., Zhong, R., Zhou, G., Leike, J., Kaplan, J., and Perez, E. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming, 2025. URL <https://arxiv.org/abs/2501.18837>.
- Stripe, Inc. Stripe identity, 2024. URL <https://stripe.com/identity>. Identity verification service with pricing starting at \$2 per verification.
- Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., Huang, S., Mountfield, A., Hong, J., Ritchie, S., Stern, M., Clarke, B., Goldberg, L., Summers, T. R., Mueller, J., McEachen, W., Mitchell, W., Carter, S., Clark, J., Kaplan, J., and Ganguli, D. Clío: Privacy-preserving insights into real-world ai use, 2024. URL <https://arxiv.org/abs/2412.13678>.
- United States Government. United states government policy for oversight of life sciences dual use research of concern, 2012. URL <https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf><https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/05/USG-Policy-for-Oversight-of-DURC-and-PEPP.pdf>.
- U.S. AI Safety Institute. Managing misuse risk for dual-use foundation models. Initial Public Draft NIST AI 800-1, U.S. AI Safety Institute, Gaithersburg, MD, July 2024. URL <https://doi.org/10.6028/NIST.AI.800-1.ipd>.
- Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., and Shi, W. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms, 2024. URL <https://arxiv.org/abs/2401.06373>.
- Zhang, S., Zhao, J., Xu, R., Feng, X., and Cui, H. Output constraints as attack surface: Exploiting structured generation to bypass llm safety mechanisms, 2025. URL <https://arxiv.org/abs/2503.24191>.

---

## A. Estimating the Number of Requests Related to Biology

To estimate how many requests in current systems are related to biology, we used the second version of the Anthropic Economic Index (Handa et al., 2025), a dataset of 1 million anonymized conversations from the Free and Pro tiers of Claude.ai.

In the dataset, the conversations are clustered by topic, and the proportion of each topic in the whole dataset is given. For example, the topic “Help with agricultural business, research, and technology projects” makes up 0.15% of the requests in the dataset. There are three levels of topic granularity; we use the lowest, most granular level.

We filtered the dataset to only include conversations whose topic contains one of the following keywords related to biology: *biolo*, *bioch*, *biotec*, *cell* (when at the beginning of the word), *genet*, *genom*, *microb*, *bacteria*, *virus*, *viral*, *proteo*, *protei*, *enzym*, *organism*, *botanic*, *zool*, *marine*, *aquat*, *anatom*, *physio*, *immune*, *neuro*, *patho*, *infect*. The total proportion of these requests was 2.98%. (For the second level of granularity, the proportion was 5.27%, and on the third level, it was 7.34%, which is likely because various science topics are subsumed under the same general topic.)

## B. Risk Tiers and Verification Levels in Bioweapon Construction

We provide a more detailed example of the access control system in the context of biology to illustrate a possible initial implementation of the system. However, we stress that this is a hypothetical example, and that an actual system would be more granular and require a lot of input from experts in the field (which we are, decidedly, not).

### B.1. Tier 0: No risk

**Verification** No verification required. This tier contains information that is already freely available in undergraduate textbooks and poses no additional risk when accessed through AI systems.

- **Standard techniques in molecular biology** including PCR amplification, DNA cloning, bacterial transformation, and gel electrophoresis. These are taught in undergraduate courses and have limited applications to building bioweapons.
- **Basic laboratory safety procedures.** These are fundamental to all biological work and widely available.

### B.2. Tier 1: Low risk

Basic ID-based verification through existing providers. This is mostly to keep an audit trail matching people to requested knowledge. This approximately corresponds to the standard BSL-1 and BSL-2.

**Verification implementation** Services like Stripe Identity provide global identity verification through government-issued ID documents, biometric verification, and address confirmation. These services already operate in most countries and cost under \$2 per verification. The integration requires standard API implementation and creates minimal operational overhead. For academic users, verification could instead utilize existing systems like ORCID (global researcher identification) or institutional email verification through services like SwiftVerify that confirm academic affiliations.

- **BSL-1 and BSL-2 organism cultivation and procedures.** While BSL-2 organisms can cause disease, they’re standard in medical laboratories and educational settings.
- **Basic CRISPR-Cas9 techniques.** These are fundamental to modern biology education and research, but could be used for dangerous modifications.
- **Large-scale fermentation and bioprocessing.** This is essential to biotechnology but could theoretically support harmful production.

### B.3. Tier 2: High risk

Maximum security for knowledge with significant weapons potential and limited legitimate applications. This approximately corresponds to the standard BSL-3 and BSL-4.

**Verification implementation** Model developers would tap into existing BSL-3/BSL-4 laboratory certification databases maintained by national authorities across major countries. In the United States, this involves accessing CDC select agent program registrations and NIH BSL-3/4 facility certification databases. In the European Union, individual member states maintain containment level (CL-3/CL-4) facility registrations that map directly to BSL levels. Canada maintains Physical Containment (PC-3/PC-4) facility databases, Australia has BSL-3/4 certification records, and Japan, China, and other major countries maintain similar systems with compatible standards. Additionally, access to high-risk knowledge might be only per-project, and each project might require approval from an expert committee.

- **BSL-3 organism cultivation.** These are biological weapons agents with limited legitimate applications.



- 
- **BSL-3 and BSL-4 laboratory procedures.** These require specialized facilities and training with security implications.
  - **Specific information about toxin genes** including botulinum, ricin, and diphtheria toxin genes, the way to obtain them, work with them, and insert them into organisms. These are extremely limited in legitimate applications.
  - **Complete viral genome synthesis techniques.** These can recreate dangerous viruses with minimal legitimate applications.
  - **Advanced CRISPR applications** including multiplex editing and epigenome editing. These have greater potential for creating dangerous modifications and are typically used only in advanced research settings.
  - **Cell membrane and surface protein modification techniques.** These are critical enablers for immune evasion.
  - **Aerosol generation techniques for biological agents.** These are specifically needed for biological weapons delivery despite some pharmaceutical applications.