Setting a High Standard for Frontier Model Security

A Policy Proposal for the Non-trivial Artificial Intelligence Grand Governance Challenge

Written by Xavi Costafreda-Fu, Beck Peterson, and Ludwig Illies

Executive Summary

Frontier AI models are 'highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety' (Anderljung et al., 2023).

Frontier model security is strong for the software industry, but not strong enough to defend against persistent, skilled, well-resourced attackers on the scale of nation-states. As frontier AI models become more capable, we expect attacks against them to scale too.

The theft or unauthorized release of a frontier AI model could threaten national security, destabilize the economy and society, and accelerate race dynamics by erasing existing competitive advantages.

In particular, if a frontier AI model is misaligned or has dangerous capabilities, then release could pose global catastrophic or existential risk, both directly and as a risk factor for other risks. We are on the precipice of such potential existential threats.

We propose the implementation of advanced security practices to raise the security baseline and to bring these risks to the forefront of the AI industry. With this preparation, we can accelerate responses when attacks become more capable, thus reducing the associated risks.

To achieve this, we argue for the development of a set of standard practices for frontier AI model security. For high-security practices, we propose the implementation of relevant security practices from high-consequence industries, such as aviation. Practices developed in these industries have been effective at protecting against hostile attacks, and much of the work in developing these practices has already been done.
Alongside this, we underscore the difficulty in communicating the need for increased security and ensuring secure implementation.

Key takeaways

- Frontier AI models are compact, easily transmitted and incredibly valuable, making them a key target for attackers.
- Yet, frontier AI models are fundamentally software making them difficult to defend.
- The breach and proliferation of an advanced frontier AI model, such as an AGI, carries significant risks. It could disrupt the economy and our society, exacerbate race dynamics between companies and between countries, threaten national security, and destabilize international geopolitics.
- Compared to similar high-stakes industries, the security standard for AI is currently much weaker.
- Our current plan, lightly held, is made up of four stages:
    a. **By 2025**: a **security survey and report for AI companies** detailing the need for greater security and making suggestions for highly beneficial practices
    b. **By 2026**: internally verified **voluntary implementation of specific advanced security practices in three or more companies.**

c. **By 2027**: industry-wide **voluntary commitment to a set of security practices** (examples outlined in the Appendix). Companies will self-regulate initially.
d. **By 2030**: transition to **government-enforced frontier model information security requirements**. The transition period between these may involve increasing government enforcement through groups such as the CISA.

We see these as goals, not actions, with instrumental steps taken over time to achieve them.

- Our policy applies only to frontier models, to target only the models with the most associated risk in the interest of increasing efficiency, as measured by our training metrics and capability bands.
- Our enforcement strategy is based on a transition from **self-regulation** to third-party/**government regulation**.
- Standards development organizations (SDOs) and US/UK federal agencies are both relevant for the top-down phase of regulation.
- Updating the policy twice yearly is suggested in order to address the most current threats.

Our proposal covers point 5, 'Information security requirements', on '12 Tentative Ideas For US AI Policy' list (Muehlhauser, 2023), and 'Security standards' and 'Military-grade information security' from 'Towards Best Practices in AI Safety and Governance: a Survey of Expert Opinion' (Schuett et al., 2023) from the AI Governance Grand Challenge brief.

Theory of Change

Status quo

Currently, leading AI companies have a security standard that is strong for the software industry. There is consistent compliance with compulsory and voluntary standards for both data protection and information security (see Microsoft, Amazon), e.g., GDPR, CCPA, and ISO/IEC 27001, and many companies have responsible disclosure and 'bug bounty' systems in place for reporting of security vulnerabilities (see OpenAI, Protect AI's Huntr). Adversarial testing, or 'red teaming', of models and security systems is also widely used.

Frontier AI model development has high risk of attack from well-resourced, persistent, malicious actors, as well as security failures leading to similar risks to national and international security and stability. For example, a frontier model could design novel biological and chemical weapons. However, compared to similarly high-stakes industries, the security standard is much weaker. Industries such as aviation and nuclear power have strict regulations to ensure their security, but no such expectation, regulatory or cultural, exists for leading AI companies.

To secure frontier AI models, all aspects of their development must be addressed, from data collection and model training to parameter protection and access monitoring. Two main considerations are the training data and the model itself. Data protection involves ensuring the integrity and security of training data by verifying the sources and contents of the data, removing any potentially dangerous data, and securely encrypting the data during transmission and storage. The model's parameters must then be protected during training and deployment. This is particularly difficult as parameters are both incredibly valuable and condensed, but also easily duplicated and transmitted.

Applying proven security practices from other high-risk industries such as two-party control used in nuclear security is a first step. We argue this approach of identifying and implementing proven practices should be applied to frontier model security across the industry.

Risks being addressed

The breach of a capable frontier model could occur in a wide range of ways and have huge negative impacts.

We break the information security risks during frontier model development and deployment down into four types: accident, system, insider, and attack risks.

- Accident risk - any unintentional leak of information through an error, malfunction, or mistake
- System risk - a persistent information leak due to flaws in system design
- Insider risk - an intentional security breach by someone with authorized access to the information

- Attack risk - A security breach by external, unauthorized actors. This includes stealing or accessing the model, or compromising the data or algorithms used to train or run the model

Note that this list is likely not comprehensive.

With access to the parameters, an attacker would be able to retrain the model for malicious purposes and potentially publish it online.
The successful destruction, theft, or publication of a frontier AI model then carries the following risks, among others:

Threats to national security

Political interference

Targeted political advertising and misinformation has already been seen in the Facebook-Cambridge Analytica data scandal and Russian interference in the 2016 United States elections. Access to a frontier model would exacerbate the problem.

Terrorism

It has been shown that models can be used in the biological and chemical weapons design and implementation process. Parameter access to frontier models could therefore facilitate terrorism on an unprecedented scale. See Anthropic's biological weapons red-teaming, and Repurposing a drug discovery model for chemical weapons development (Urbina et al., 2022).

Geopolitical implications

International access to the most advanced frontier AI models would lose the USA its competitive advantage, destabilizing international politics and trade.
Frontier model theft could also come from hostile state actors, raising tensions, deepening divisions, and increasing the risk of great power conflict, even if the source of the breach was uncertain.

Rapid economic instability and societal disruption

Mass proliferation of an advanced frontier model could allow for simultaneous replacement of human labor and huge economic growth. If done too quickly, this could destabilize the economy and markets. This would lead to huge disruption to the status quo, with possibilities of rioting or revolt and dangerous concentration of profits among select corporations that exploited the shift.

Intensified race dynamics

Universal access to a private, advanced frontier model would eliminate any competitive advantage through intellectual property. Companies would push to progress more rapidly, likely at the expense of safety and security precautions. This would accelerate capabilities timelines and increase competition:
a) Between companies, and
b) Between countries.

a) Between companies

The uncontrolled proliferation of a cutting-edge frontier model would erase existing competitive advantages based on intellectual property. This loses industry leaders their advantage and gives less successful companies the opportunity to get ahead. Advantages in talent and compute would remain, but the open-source community would make the breached model available for free.
All AI companies would be incentivised to rapidly develop the capabilities of their models in order to maintain profits and stay competitive. This could come at the

expense of AI safety and security, particularly from companies without a talent or compute advantage, increasing the risk of further breaches and potentially catastrophic misalignment.

b) **Between countries**
The same dynamic applies internationally. Trailing countries would be incentivised to invest in their model development programs, accelerating timelines, endangering international cooperation, and hindering safety efforts.

How will policymakers measure efficacy?

The security industry already has a strong precedent of monitoring and reporting attacks and vulnerabilities; ISO/IEC 27001:2022, the main international information security standard, contains 'Section 9 - performance evaluation' (see page 4).
The policy can be assessed through these existing systems, with measurement of efficacy through both attack reports and general incident and issue reporting.

Specifically, we believe confidential sharing of attacks on frontier AI models with policymakers would help evaluate the efficacy of the policy. Sharing would include detailed information on attacks that were fully stopped, partially stopped, and not stopped to give a full understanding to policymakers.

How can it be adjusted to fit changing trends?

See '*Updating the policy*'

Key Design Parameters

Who the policy applies to

The policy applies to any company developing frontier AI models, under the Anderljung et al. (2023) definition of 'highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety'.
Currently, this consists of six groups: Amazon, Anthropic, Google (including Google DeepMind), Meta, Microsoft, and OpenAI. These companies have the most well-known and capable models and are key targets for well-resourced malicious attackers. Except for Google DeepMind in London, UK, the current group is concentrated in only two states of the USA, California and Washington, making full coverage by the policy easier. We suggest focusing on establishing US regulation first, and then expanding to other countries.

As more resources are allocated to frontier model development, **we expect the number of companies within the policy's scope to increase**. Improvements in training algorithms will lower the cost of training models with today's cutting-edge capabilities, e.g., GPT-4, Claude, DALL-E. We predict this will be balanced by continued compute scaling and the cost of frontier model training will continue to increase over the next 5 years. However, this will likely be outpaced by increased investment, leading to more companies joining frontier model development in the next 5 years and coming within the policy's scope.

In which situations the policy applies

Within a given company, **we suggest only applying the policy to systems and people involved in frontier model development and deployment**. This limits the costs and inconvenience of new security practices and concentrates the security benefits to where they are most needed.

Included systems and personnel would include:

- Compute clusters for training the models
- Training data for the models
- Frontier model parameters
- Technical employees working on frontier AI models
- Non-technical employees connected to frontier AI models

How the policy will be enforced

Following our two-stage implementation, enforcement will come from two groups:

1. **In the first stage, we propose companies enforce the policies internally**, as we aim to ride the current push for collaboration and safety. The downside of this is that companies will self-regulate any commitments, likely leading to partial or ineffective adoption from some. This compromise forms a placeholder solution for accelerated adoption.
2. **In the second stage, the policy will be government-regulated** by either a new US regulatory body for AI, if one is created, or a state department such as the Cybersecurity and Infrastructure Security Agency (CISA) who already have experience conducting security assessments.

Enforcement is covered in more depth in the next section, *Comparing Implementations*.

Who will verify the policy's implementation

Different groups will verify effective implementation as regulation transfers from self-regulation to governmental regulation through the hybrid transition period. **Initially, companies will verify implementation through existing internal auditing and reporting systems, and regular third-party compliance reviews**. Both methods are already in use in large software companies to review information security management systems, so any security practices implemented under the policy would likely be automatically included.

After initial industry adoption, we recommend incorporating frontier model security evaluations into existing state systems, such as the CISA's review system, as they already have the experience and knowledge to make accurate assessments of high-security sectors. This is from a shallow investigation, however; we have not investigated the efficacy of these assessments, and there are likely better solutions.

Timescales

Frontier model capabilities are rapidly improving along with investment in the sector. This makes them very high value targets, as outlined in 'Theory of Change', and a rapid, unforeseen improvement in capabilities could lead to a rapid increase in attacks while security systems are still unprepared.
Therefore, we have to pre-empt these potential disjoints through quick, effective implementation. **We aim for adoption of advanced security practices in three companies by the end of 2025, unanimous frontier model industry commitment by 2027, and government regulation and accountability by 2030. We recognise that this timeline is uncertain**. There would be significant overlap between these goals, but these markers will be updated as we learn more about the object-level state of AI governance.

How often the policy will be updated

Considering how rapidly the frontier model industry moves, **we recommend updating the policy and set of standard security practices twice-yearly**. The policy aims to preempt future advances in attacker capabilities, which will likely move with the frontier model industry, so a regular, short review cycle to add, remove and update the standard is essential.
Reviewing the proposal is covered in more depth in 'Updating the Policy'

How long the policy will last for

We expect our involvement in this policy to last for less than five years (80% certainty as a ballpark), and likely less than two years (60%), as we either fail to get it implemented or it gets taken over by the industry or government.
That being said, there is no end date for high security unless there is a fundamental change in the ongoing offense-defense arms race. This policy could continue to be updated for decades, as it provides a general system for improvements.

Comparing Implementations

Security can be improved through suggestions, commitments, and standards, with each meeting varying levels of precision. These can either be trust-based, relying on cooperation, or enforced by a regulatory body, either private or public.

A standard is best for securing frontier model development as it can enforce specific security strategies, such as zero-trust architecture, and provides a clear security requirement. The recent White House Voluntary Commitments include undetailed but high-level security assurances, so a concrete, actionable standard to implement these commitments is the next step.

Enforcement

The standard could be enforced by the government, a private third-party, or the companies themselves. Based on our theory of change, we will approach companies first, as that seems to be the fastest path to impact.

Government enforcement is most effective as penalties can be used to ensure cooperation, but may face bureaucratic and political obstacles and take longer to implement. NIST is the key public standards development organization (SDO) in the United States, and groups such as the Nuclear Regulatory Commission and Cybersecurity & Infrastructure Security Agency (CISA) conduct assessments of high-security facilities.

Private organizations, either for-profit or non-profit, circumvent these difficulties and can adapt faster, making them better suited for quick implementation. The incentives of these groups, however, can be misaligned with securing frontier AI models, e.g. if a non-profit organization relies on donations from the companies it is regulating. The IETF is a good example of an effective non-profit SDO that establishes standards for the Internet.

Self-enforcement by companies, or mutual enforcement between them, do not require a regulatory body and are therefore faster to implement and adapt. One existing method is a self-reporting system, which relies on company reputation and industry expectations of cautious openness. However, the companies are not impartial here and may have conflicting interests, leading to weak enforcement. Existing groups, such as the Frontier Model Forum, are yet to produce results as of August 2023, but effective industry-led groups are possible: the PCI Security Standards Council is a strong example of standard creation and self-enforced adoption from the payments industry.

Coordination

The security standard must be widely adopted long-term. There are several leading companies in model development and a breach in any advanced system is dangerous, so all frontier systems must be secured.
Self-regulation is not sufficient long-term, as companies can act unilaterally and invest less money into security in order to gain a competitive advantage. On the other hand, it requires less coordination to implement practices, so it is still a strong first step.

For practical implementation, we plan for adoption from one company at a time to ease the process and make improvements. However, this policy could be adopted by all 7 leading companies at once under '*3 - Invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights*' from the Voluntary Commitments.

Our Recommendation

A high-precision standard that is externally enforced would best ensure consistent security across the industry. However, this would require a regulatory organization, either public or private, increasing time to implementation.

Considering the urgency of securing frontier AI models, we suggest a leveled four-stage implementation:

1. **By 2025**: a **security survey and report for AI companies** detailing the need for greater security and making suggestions for highly beneficial practices
2. **By 2026**: internally verified **voluntary implementation of specific advanced security practices in three companies**
3. **By 2027**: industry-wide **voluntary commitment to a specific set of security practices** (examples outlined in the Appendix). Companies will self-regulate initially.
4. **By 2030**: an industry transition to **government-enforced frontier model information security requirements**, once the necessary regulatory bodies have been established. The transition period between these may involve increasing government enforcement through groups such as the CISA.

Technical Features of AI

Computation metrics

The AI Triad of compute, algorithms, and training data, combined with model size measured in number of parameters, provides the key technical metrics for evaluating the resources used to train a frontier model.

The two points relevant to our policy are:

1. Language model capabilities scale with model size, dataset size, and compute used in training (Kaplan et al., 2020).
2. "every 9 months, the introduction of better algorithms [to computer vision models] contribute the equivalent of a doubling of compute budgets", ('Revisiting Algorithmic Progress'; see Figure 1)

Figure 1: a breakdown of computer vision model improvements over time according to the AI Triad. Note the significant contribution of algorithmic progress ('Revisiting Algorithmic Progress')

	Reduction in error	Algorithmic progress	Compute scaling	Data scaling
AlexNet → ResNet50	23.7	64.9%	35.1%	NS
AlexNet → ResNeXt-101	24.0	70.6%	29.3%	NS
AlexNet → BiT-L	24.2	40.8%	47.2%	12.1%
AlexNet → VIT-H/14	24.8	43.7%	44.4%	11.9%
AlexNet → VIT-e	27.6	41.6%	43.6%	14.8%
ResNet50 → BiT-L	10.4	30.7%	47.3%	22.0%
ResNet50 → VIT-H/14	10.9	35.2%	43.4%	21.4%
ResNet50 → VIT-e	13.8	34.1%	40.9%	25.0%
ResNeXt-101 → BiT-L	6.6	24.9%	49.8%	25.4%
ResNeXt-101 → VIT-H/14	7.2	30.1%	45.3%	24.5%
ResNeXt-101 → VIT-e	10.0	30.3%	41.6%	28.1%

With this in mind, we suggest using the following technical metrics to measure model size:

- Floating point operations (FLOPs) to measure compute used in training, calculated through either:
    - Combining information about the architecture and the dataset
    - Using information about the hardware used and the training time.

  See 'Estimating Training Compute of Deep Learning Models' for more detail on calculations
- Number of parameters, or size of the trained model in gigabytes (GB)
- Training dataset size in gigabytes (GB) to measure the dataset
- Qualitative consideration of significant algorithmic improvements to account for algorithmic progress
    - e.g., from sigmoid to ReLU activation functions

Due to significant uncertainty over confidential details such as the specific architecture used, these metrics can only be used as an estimation for the range of a model's potential capabilities.

Capability bands of models

We separate current capabilities into two bands: frontier / foundation models, and narrow models.

We use the Anderljung et al. (2023) definition of frontier AI models:
'highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety'

And the Ada Lovelace Institute's definition of foundation models:
'Foundation models are AI models designed to produce a wide and general variety of outputs. They are capable of a range of possible tasks and applications, such as text, image or audio generation. They can be standalone systems or can be used as a 'base' for many other applications.'

We define Narrow AI as:
'Any model trained to perform a specific task with few, or no, wider applications'.

Our policy would cover only frontier model security to reduce costs and inconvenience while still securing the most potentially dangerous models.
Narrow AI, such as Inflection's Pi, are considered out of scope and would not be included at this stage.

Capability evaluations will be key in determining the potential risk of a frontier model if breached, with groups like ARC Evals already doing promising work on self-replication in this area. We anticipate the development of more extensive evaluation systems in the next few years, and will use these as they are created to judge the risk of a given frontier model.

Within frontier AI models, we suggest using company valuation for comparisons of model value. Company valuation represents the value the company provides to the public, and, as a proxy, the value of their model. Companies with higher valuations have higher-value models, and will thus be the targets of more attacks of higher sophistication. Public valuation and investment will be most useful when comparing between frontier model companies as there may be narrow AI companies, such as Inflection, which have capital from investment but whose models do not pose a significant threat if breached.

Relevant policymakers

We think the most relevant policymakers differ for the two phases of the policy, with decision-makers in companies most important for bottom-up governance, and Congress most important for top-down regulation.

Individual Companies

Our policy can be implemented on a company-by-company basis. This is more targeted and eliminates coordination problems, but full coverage is slower than multilateral industry adoption.
The speed and precision of individual adoption is well-suited to our first stage of adoption of advanced practices by the end of 2025. With this, we're aiming to expand the Overton window for which security practices are standard or reasonable to make later industry adoption easier.

Companies

Frontier Model Forum

The Frontier Model Forum, currently Anthropic, Google, Microsoft, and OpenAI, is "*an industry body focused on ensuring safe and responsible development of frontier AI models*". Two of its four core objectives are relevant to frontier model security:

1. "*Advancing AI safety research to promote responsible development of frontier AI models, minimize risks, and enable independent, standardized evaluations of capabilities and safety.*
2. ***Identifying best practices for the responsible development and deployment of frontier AI models***"

Our proposal turns the second objective into a concrete, actionable policy that can be adopted by the Frontier Model Forum and, in turn, four of the major companies in frontier model development

National government

The United States

The White House has shown support for responsible private AI development with the White House Voluntary AI Commitments, and also for improving cybersecurity in federal systems and critical infrastructure with a National Security Memorandum on cybersecurity in critical infrastructure, and Executive Order 14028. These concrete outputs are promising, but the Voluntary Commitments already cover frontier model security in vague terms, and the White House is likely most relevant for publicity and gaining support for regulation, not for specific, detailed legislation.
Congress also holds significant power and has already expressed interest in AI governance. The United States Senate Judiciary Subcommittee on Privacy, Technology and the Law has already had two hearings from CEOs, academics, and think tank leaders (May, quotes; July, partial transcript) with a focus on the risks from AI and potential regulation. These hearings

have not produced any concrete initiatives yet, but they demonstrate Congress' current bipartisan recognition of the risks posed by AI. As a result, **we see Congress as a promising policymaker long-term for implementing fundamental legislation on responsible AI development and deployment**.

The United Kingdom

Although most frontier AI models are currently located in the US, Britain developed a national AI strategy in December 2022. These efforts by Britain to increase its influence within the AI sector and set "an example in the safe and ethical deployment of AI" makes them worthy of consideration.
The Foundation Models Taskforce, and the global AI summit at Bletchley Park in November are the two most promising outputs of this strategy to date.

Our policy could complement the statement, " The UK public sector will lead the way by setting an example for the safe and ethical deployment of AI through how it governs its own use of the technology", by providing a framework for company self-regulation. The UK also created the AI standards hub in 2022, although its output is limited so far.

Existing Governance Structures

Other industries with a similar policy

Frontier model security is unprecedented in its difficulty, but not its class of attackers. Both public and private industries have been the target of well-resourced, persistent, and knowledgeable malicious actors: nuclear power and weapons, aviation, and finance are three key examples.
The security systems around weapons of mass destruction and nuclear power are particularly relevant as they include concrete information security, physical security, personnel security, and cybersecurity practices that are applicable to frontier model security.
These include:

- Two-party control
- Insider threat programs
- Air-gapping
- Safety and security by design
- Building a security culture

Why can these policies be applied to AI?

The theft of a highly capable frontier model carries destabilization and disruption risks through escalation of race dynamics and misuse risk in a similar way to weapons of mass destruction or nuclear fuel. These risks must be mitigated, so similar approaches are called for.

Expanding on existing laws and regulations

We aim to build off the precedent of the White House Voluntary Commitments, signed by seven leading AI companies. All major companies developing frontier AI models have signed these Voluntary Commitments. They include:

- "Invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights"

Our policy is a concrete way to fulfill this commitment, which may make it more appealing to these companies. The White House is also drafting an Executive Order to follow up on the Voluntary Commitments.

Mandates of existing departments / public offices

Standards development organizations (SDOs) and US federal agencies are both relevant for the top-down phase of regulation.
The **National Institute of Standards and Technology (NIST)** and the **International Organization for Standardisation (ISO)** are the most relevant SDOs. These two groups develop key information security and cybersecurity standards, such as the Cybersecurity Framework 2.0 and SP-800 Series from NIST, and ISO/IEC 27001:2022 from the ISO and International Electrotechnical Commission.
While these groups are not well-suited to quick development and implementation, they may, however, be useful for formalizing the standard practices long-term.

The Department of Homeland Security, Cybersecurity and Infrastructure Security Agency, and the National Security Agency are the most relevant federal agencies for our policy, since they already have experience with securing critical systems. Within these, **we see getting machine learning models classified as critical infrastructure as part of the Information Technology Sector as an important long-term goal** to ensure public funding of security measures and safety research**.** This may have downside risk, and we have not done a deep investigation into the potential implications of classing AI as critical infrastructure.

Updating the Policy

Updating the policy while maintaining security is the key difficulty here. A precautionary approach here is needed as trial and error is not an option with such high stakes. We suggest preemptively implementing high-security practices, and removing any that have been shown to be ineffective through a bi-annual review system.

We expect containment of frontier AI models to become much more difficult as models become more powerful for several reasons:

1. Attackers will spend more resources on trying to steal, destroy, or disrupt frontier AI models as their capabilities and value increase.
2. Publicly available and privately developed AI models will likely become better at offensive cyber attacks, reducing the costs of an effective attack and reducing barriers to entry for attackers.
3. Models may not be fully aligned and attempt to break out of the system and self-replicate through the Internet.

These will all require more stringent security measures. Examples include:

- Air-gapping: full isolation of any system involved in frontier model development. This could potentially apply to deployment too, though further research is needed to determine whether this is necessary.
- Escalating insider threat programs: moving from basic personnel screening to rigorous, regular review systems, e.g., the US military's Nuclear Weapons Personnel Reliability Program
    - "*Only those individuals who demonstrate the highest levels of integrity and dependability will be chosen for PRP duties.*" - from the PRP manual

We suggest a regular twice-yearly review system to implement these additions, on top of the option to make emergency updates if unknown vulnerabilities are exploited by attackers. **These changes may not be publicly published; this gives companies time to implement policies without publicizing existing vulnerabilities**.

These reviews would also include improving on or removing practices that have been shown to be ineffective, either through a successful attack through or around the practice, or a demonstrated technical vulnerability that has not yet been exploited, e.g., a known zero-day vulnerability.

Some relevant areas for evaluating a practice include: coverage; security benefits; usability; time to implement; cost to implement; running costs / reliability; compatibility with existing systems; scalability; relevant metrics / transparency; adaptability and resilience; and any historical attacks against the practice, both failed and successful.