Introduction

Learning about cryptography, in particular, about privacy-preserving techniques such as Secure Multi-Party Computation and Fully Homomorphic Encryption over the past few months I realised the potential applicability of these techniques in the problem of AI Boxing. This led me to explore the pragmatic side of AI boxing, the issue of centralisation of power if cryptographic boxing were to be implemented, and subsequently what could be done about it.

In this post, I will primarily build on the ideas covered in this post, which I encourage you to read if you haven't. The summary of what I am going to build on is as follows:

- Boxing will likely become *necessary* relatively soon, as we are making little progress on alignment relative to progress being made in AI capabilities.
- For sufficiently capable AI models it is better (from the risk reduction perspective) to remove any information channels (even in the form of interpretability tools) than to add them.
- Social engineering is likely going to be the biggest failure mode of AI boxing.
- More efforts should be employed in optimising boxing techniques and we should develop a better understanding of what information channels to read and when.

And with that in context I am going to argue that:

- While boxing becomes necessary at a certain AI capability threshold, in practice we are unlikely to implement it due to very high costs and economic disincentives.
- Under the assumption that boxing techniques were to be implemented despite the enormous alignment tax they bear, our best ideas/schemes for boxing would lead to the centralization of power. This centralisation of power could have adverse effects, bearing risks comparable to those of not implementing any containment in the first place.

Lastly, I am going to propose and briefly discuss a scheme that could, if implemented, mitigate the issues associated with the centralization of power created by "standard" boxing schemes.

Quick note: I am very far from an expert in cryptography, cyber-security, or alignment, therefore all feedback would be greatly appreciated!

Boxing (Re)Definition

AI Boxing is generally defined as the problem of ensuring that AI is securely contained. That is, it cannot interact with the outside world. The problem also concerns preventing social engineering through the remaining available information channels. However, this frames the problem in a more theoretical than pragmatic way, neglecting the real-world viability of potential solutions. Theoretical boxing schemes are useful as they allow us to explore what is possible, and I am going to propose one later in this post. However, most of them won't be implemented due to the astronomically high alignment tax they bear. So while knowing what is possible is useful, I believe that now it is more important to know what is feasible.

The Alignment Tax of Boxing AI

The costs of boxing vary as the ways of constraining the AI's ability to access the world outside of containment do. The cost of implementing a given scheme is roughly proportional to how much it constrains AI access. I am going to split these costs into three categories:

1. Computational overhead inherent in boxing schemes. (mostly due to the cryptographical techniques used)
2. The opportunity cost of limiting AI access.
3. The physical costs of building a new, safe, infrastructure.

1. Computation overhead inherent to boxing schemes

Two cryptographic techniques likely to be used in any boxing scheme are Fully Homomorphic Encryption (FHE) and, potentially in the case of decentralising computation, Secure Multi-Party Computation (MPC). More techniques such as zero-knowledge proofs could be used in various schemes. However, I will focus mainly on FHE and MPC in this post.

Fully Homomorphic Encryption (FHE) allows for computations on encrypted data without decrypting it. This is usually done in the following way:

1. Key Generation
2. Key Sharing (this is an optional step, only used when encryption is shared by multiple parties)
3. Encryption
4. Homomorphic Operations (addition and multiplication of ciphertexts)
5. Decryption

FHE has a problem with noise accumulation during homomorphic operations which causes a major performance problem. This is mitigated by a noise budget, which when exceeded causes decryption to fail. This in turn is solved by "Bootstrapping", a technique that refreshes ciphertexts while maintaining their homomorphic properties. However, this operation is heavily computationally expensive. The current best bootstrapping techniques were able to reduce the number of homomorphic multiplications to $O(log^2\lambda)$, where λ is the security parameter, which is lower than the previous SOTA performance of $\tilde{O}(\lambda^4)$. This however is still incredibly computationally expensive. The time complexity and computational overhead of each step of performing FHE is detailed below:

Operation	**Time Complexity**	**Computational Overhead**
Key Generation	O(λ^3) to O(λ^4)	Very high overhead. One-time cost in many protocols.
Encryption	O(m * λ) to O(λ^4)	High overhead. m is the message length. Results in significant ciphertext expansion.
Homomorphic Addition	O(n)	Moderate overhead compared to other FHE operations, but still significantly slower than plaintext addition as the operation is done on ciphertext instead of plaintext.
Homomorphic Multiplication	O(n log n) to O(n^2)	Very high overhead. Significantly slower than plaintext multiplication. Complexity depends on the specific FHE scheme.
Noise Management (Bootstrapping)	$O(log^2\lambda)$ or higher	Very high overhead. Critical for maintaining correctness during computations. Substantially increases overall computational cost.
Decryption	O(m * λ)	Moderate to high overhead. m is the message length. Generally less intensive than encryption but still significant.
Overall	O(λ^4 * 2^L)	Where L is the circuit depth. A large part of the exponential complexity comes from bootstrapping.

Another way to estimate the computational overhead is to look at the efficiency of existing optimised schemes and use that as an indication for what could be expected from schemes in the future. A SOTA FHE scheme that runs a CNN on MNIST takes 1.8 seconds to

perform an evaluation pass on 1 image. On my "potato" laptop, I evaluated all 10,000 images in the dataset in 9.1 seconds, that is 0.00091 seconds per image (or 0.91 milliseconds). This roughly indicates a 1,978x increase in time per image evaluation between the optimised FHE scheme and baseline (my laptop). This comparison is far from perfect but it gives us another clue to the magnitude of the overhead that we can expect.

My estimated interval for computational overhead for any FHE scheme implemented for AI boxing is:

- Lower bound: 100x more computation.
- Upper bound: 1,000,000x more computation.

That is to say, even with algorithmic improvements and hardware optimisations, we are likely looking at at least **100x computational overhead** by using FHE.

Secure Multi-Party Computation (MPC) allows for multiple parties to jointly compute a function over their inputs without making these inputs known to other parties. In the context of AI boxing, it could allow us to perform inference without revealing the model weights or the computation being performed to a single party involved. MPC usually works in the following way:

1. **Input sharing**: Each party splits their input into shares and distributes them among all parties. This is often done using secret sharing schemes like Shamir's Secret Sharing.
2. **Circuit representation**: The function to be computed is represented as a boolean or arithmetic circuit.
3. **Gate-by-gate computation**: Parties perform local computations on their shares for each gate in the circuit.
4. **Communication**: Parties exchange intermediate results (still in shared form).
5. **Output reconstruction**: The final result is reconstructed from the shares held by all parties.

The main computational overhead in MPC comes from two things:

1. Communication between parties. This is the main bottleneck, as all data must be split and shared between all parties for each calculation step (i.e. layer forward pass), it is essentially using the network as the bus for computation. Even a fast, low-latency network, would greatly bottleneck the computation.
2. Cryptographic operations (e.g., encryption, decryption, secret sharing). These often include FHE and zero-knowledge proofs (ZK proofs) which are both computationally expensive.

The general time complexity and computational overhead for MPC is detailed below:

Component	**Time Complexity**	**Computational Overhead**
Communication	O(r * n * m) r: rounds n: parties m: message size	Network latency: ms to s per round Bandwidth: KB to GB depending on protocol and input size
Secret Sharing	O(n) for n parties	Relatively light, typically ms
Homomorphic Encryption	O(λ^3) or higher λ: security parameter	Very high, seconds to minutes per operation.
Oblivious Transfer	O(λ) per transfer	Moderate, ms to s
Zero-Knowledge Proofs	O(|C|) to O(|C| log |C|) |C|: circuit size	Substantial, s to min for complex statements
Overall MPC Protocol	O(|C|) to O(n^4) Varies widely based on the specific protocol	Highly variable, from seconds for simple computations to hours for complex ones with many parties

Unfortunately, determining the exact computational overhead for FHE or MPC is impossible as it varies heavily depending on the *exact* scheme employed. The number of parties involved, the exact type of computation performed, hardware optimisation or specialisation, bootstrapping and in the case of MPC network speed and latency play a major role in determining the efficiency of a given scheme. However, even without any precise estimates, we know that both techniques, especially FHE, *come with a massive computational overhead*[1]. The amount of work being done to make these schemes efficient enough to reach a level of usability in commercial applications is also a good indicator of just how expensive they are - they are still mostly at the “yet to reach real-time services maturity” level.

In summary, estimating the computational overhead for FHE and MPC is difficult as exact implementation specification matters a lot. However, we can reasonably expect the computational overhead of these techniques to be at least 2 or 4 orders of magnitude that of standard non-encrypted computation, or more.

[1] Direct quote from Practical Solutions in Fully Homomorphic Encryption - A Survey Analyzing Existing Acceleration Methods

2. Opportunity cost of limiting AI access

As recent history shows, the implicit consensus in how general-purpose AI development should be approached changed from a cautious one ("AI must not 'move fast and break things'")[2] to a more aggressive and profit-driven one (especially under the motivation to deploy AI products at scale). While the conservative and careful approaches did not get abandoned[3], the current race dynamics, and general hype around anything LLM or AI-related in combination with the unchangingly large benefits to be had from deploying general-purpose AI make the cost of restricting AI access extremely high. If the primary motivation for most work on AI capabilities is to deploy or utilise them for profit[4], then even the slightest restrictions on model accessibility would go against that very motivation. I don't know what the primary motivation behind creating ever more powerful AI is but any investment is expected to bring a return (especially if it is in billions of dollars). The general capitalistic notion of "maximising shareholder value" generally disincentives any form of boxing schemes. Of course, at some point, the companies/AI Labs are going to decide that the deployment or even training/testing of a certain AI model might be too dangerous to be done without costly safeguards in place. However, considering the enormous expense associated with training SOTA models, the level of risk would need to be extremely high to justify the high costs associated with implementing cryptographic safeguards. It is hard to say at what level of risk AI labs will decide that the costs are worth paying, but considering how massive the computational overhead is, it is likely going to be too late.

An additional aspect worth considering is that if the cost of running a future powerful AI, *i.e. something equivalent to GPT-6*, remains as cheap as they are now, we might be incentivised to deploy future powerful models serendipitously as opposed to efficiently. That is to say, the potential benefit to be had from the AI finding a novel or better solution to a problem it was deployed to solve is offset by the increase in the Δ deployment cost (the difference between the cost of deploying a "capable-enough" AI and more capable than necessary or SOTA AI). So, as opposed to deploying models with X level of capability for tasks that require no more than X capabilities, if the cost increase of deploying a more powerful model is not much greater, we are incentivised to deploy AI with capabilities greater than those required to accomplish a given task. So, if model deployment is cheap, we can expect this type of deployment to be desired by relevant actors, further increasing the (opportunity) cost of restricting access to AI.

---

[2] A direct quote from Demis Hassabis
[3] Especially considering the effect of the "AI race" on the public awareness of AI Safety.
[4] While there are organisations motivated by the long-term benefits to the society from creating AGI or SI, they are likely to be a lot more safety focused (at least I hope so).

In summary, deployment might be the primary motivation behind the development of powerful AI in the first place, the opportunity cost of *not deploying* is likely a considerable counterforce against implementing boxing schemes.

3. The physical costs of building a new and safe infrastructure.

Lastly, the costs associated with the construction of new infrastructure should be considered as a factor disincentivizing boxing. However, these become only relevant if both (1) computational overhead and (2) opportunity costs of limiting AI access are paid. This means that costs associated with (1) and (2) were either decreased substantially, or the safety budget permitted them. In either case, I'd expect the cost of constructing new infrastructure, however big, to be a less concerning factor, as it is mostly "fixed" and likely a fraction of (1) and (2). An example of the physical costs is the construction of an air-gapped facility housing a mainframe, with electromagnetic radiation shielding[5] and a high degree of security. Such a facility would cost a substantial amount, however as already mentioned above, likely a fraction of the costs associated with (1) and (2).

However, an important factor to consider here is the time it would take to build such infrastructure. In a market where timelines are short and even 6 months makes a substantial difference, waiting more than a year can be viewed as a big "cost" to be paid.

Centralisation of Power

Assuming that the AI Labs are willing to sacrifice the economic gains, user feedback and all other benefits to be had from deploying powerful AI, in order to decrease the risks posed by powerful misaligned AI. AI with capabilities exceeding that of an average human would likely pose risks significant enough for us to seriously worry about all *edge cases* of containment schemes, such as electromagnetic wave communication, social engineering, or any information leakage outside the containment[6]. This implies that probably the safest way to contain such AI would involve an air-gapped facility, with plenty of physical safety layers. This, by default, leaves us in a situation where only the labs (and/or potentially the government) have access to powerful AI. Thus any boxing scheme, when implemented, comes with one major flaw, *who* has access to the AI?

[5] I don't know if "simply" wrapping a facility in a Faraday cage or locating it underground would suffice, but I am quite certain that electromagnetic radiation shielding should be a rather straightforward problem to solve.
[6] It seems to me that there is almost no limit to how paranoid one can get in this regard. In example it is possible for the AI to attempt to communicate via changes in usage of its cooling system.

When I think of this kind of setup, the Manhattan Project immediately comes to mind[7], but the analogy is not perfect. I believe that it is safe to assume that all world leaders understand the significance of the impact AI will have on global power dynamics[8], but whether this awareness leads to collaboration or a cold war, is not for me to speculate on. If boxing of this kind gets implemented by one organisation/nation, we will find ourselves in a place where the greatest power accessible to humanity in history is in the hands of few.

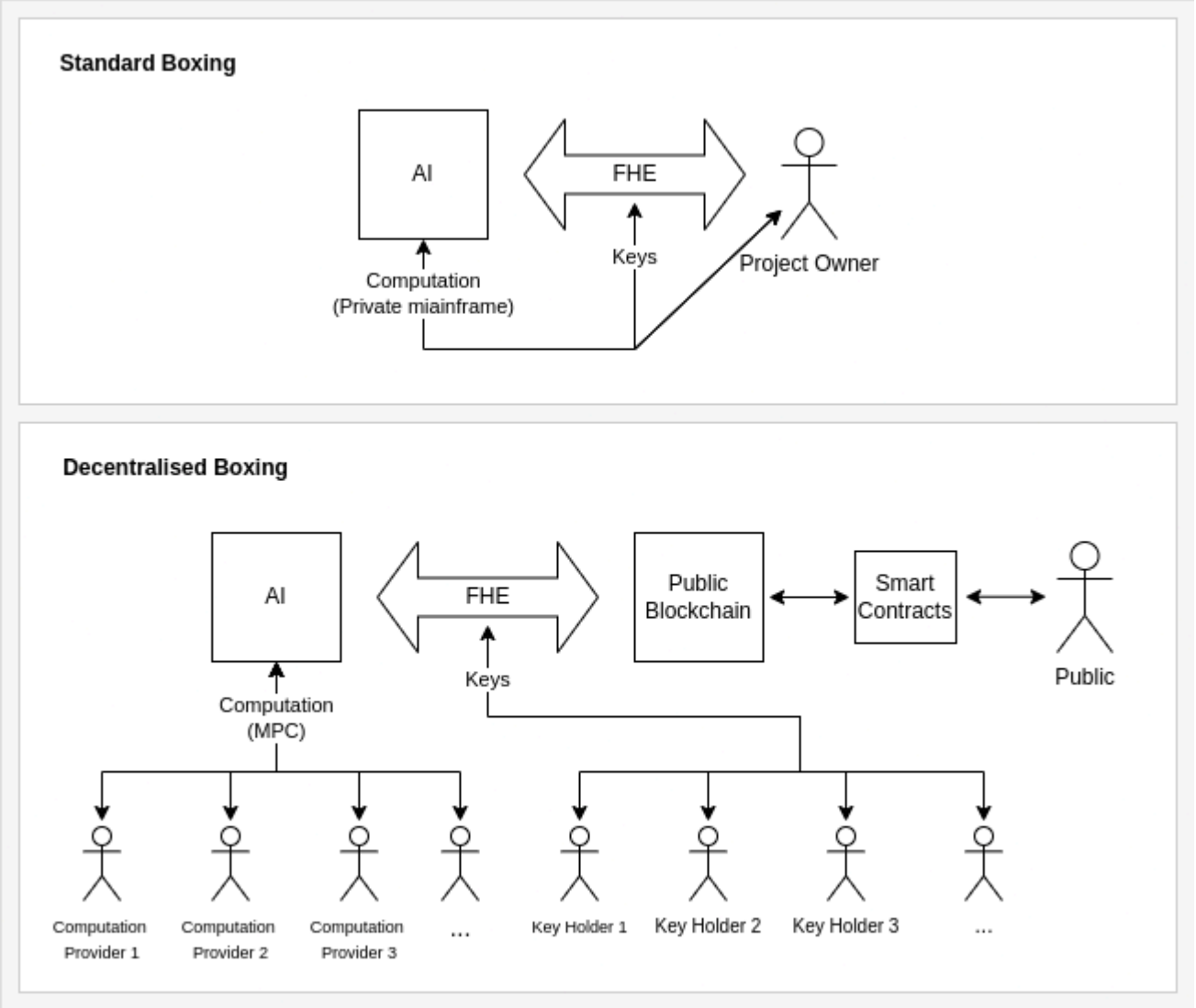
Decentralised Boxing

If boxing is unlikely to be implemented, decentralised boxing is even less likely to be implemented, primarily for two reasons 1) Additional computational overhead introduced by MPC or other distributed computing schemes, and 2) it would go against a substantial part of the motivation to build it in the first place, which is to gain power. With that said, it is in principle possible, though I don't know how feasible, to box AI in the following scheme:

Simply combining MPC and FHE so the computation and key shares (for model output decryption) can be split among all parties involved does not fully solve the problem. During the setup all parties need to come together and agree on the scheme, in other words, a form of coalition is required. Instead of a single entity, we now have a party, or a collaboration between different parties who can access the AI, so the problem is only somewhat mitigated. This type of scheme could be sufficient to prevent a substantial part of the potential misuse, as to make any change in the system a majority vote would be required from all parties involved.

Splitting the keys between a group of parties could reduce the potential misuse, but would still leave room for collusion. A way in which collusion could be mitigated is to make the information channel at least partly public (i.e. by using a public, cryptographically secure information channel). This would however greatly reduce the security of the system, as all outputs would then necessarily need to be made public. This leaves us in a dilemma between risks from centralisation/collusion and risks from misaligned AI potentially communicating with the outside world.

[7]A there is no world war, and the project motivations are likely different. However, the motivation behind the technological race to build AGI to gain technological supremacy could be viewed as comparable to that behind Manhattan Project during the second world war.
[8]I recall Vladimir Putin saying "It [AI] comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world." back in 2017.



Example of a potential implementation of a decentralised boxing scheme.

In this decentralised boxing set-up, where computation is done using Secure MPC, AI is within the FHE scheme and the encryption keys are split among trusted parties, requiring a majority vote to decrypt output, which then gets sent to a cryptographically validated public input/output channel, we have a lot of room to play with specifying who exactly are the parties providing computation, parties holding keys (could be the same ones as providing computation), and how much voting rights does the public hold. The computation could be done by only a handful of trusted organisations which could permit corruption or collusion, or it could be a lot more decentralised which is likely to increase the computational overhead. The key shares required for decryption (or any changes done with the set-up) could be split almost arbitrarily (though it gets more computationally expensive with more parties involved). The threshold at which the decryption is valid is also a subject of choice. Lastly, the degree to which the public is informed is also subject to choosing, with a potential option where the input-output channel is cryptographically bound to be public, i.e. using a public blockchain. In the case of the example above, if smart contracts are used, any form of logic or algorithm could be

used to govern the system. Of course, this all depends on the robustness of the blockchain it is built on.

In conclusion, modern cryptography could prevent the centralisation of power caused by boxing. However, this should serve more as an illustrative example of what could be done than a robust plan. On the pragmatic side of boxing, especially decentralised boxing, computational overhead remains too high to be considered as a viable solution in the near future.

Key Points and Summary

1. Boxing may become necessary soon due to the rapid progress in AI capabilities outpacing alignment efforts.
2. Implementing boxing techniques faces significant challenges:
    - High computational overhead due to cryptographic methods like Fully Homomorphic Encryption (FHE) and Secure Multi-Party Computation (MPC).
    - Substantial opportunity costs from limiting AI access and deployment.
    - Physical costs of building secure infrastructure.
3. The estimated computational overhead for FHE schemes ranges from 100x to 1,000,000x more than standard computation.
4. Current market dynamics and profit motives create strong disincentives for implementing boxing schemes.
5. If implemented, boxing could lead to centralisation of power, with only select organisations or governments having access to powerful AI.
6. A decentralised boxing scheme using MPC and FHE was outlined, showing how distributed computation and access among multiple parties could be possible, potentially mitigating power centralization issues.
7. However, decentralised boxing faces even greater implementation challenges due to additional computational overhead and conflicting interests.
8. Modern cryptography could theoretically prevent power centralization caused by boxing, but practical implementation remains infeasible due to high computational costs.

This is cool but has no place in the post:

AGI Roomba

While I don't expect "AGI Roomba" to be the end of humanity, businesses and corporations by default are generally incentivised to deploy more aggressive and risky automation strategies to gain a competitive edge. Without any restrictions on who has access to powerful AI, this quickly leads us to a place where not using aggressive automation leaves you disadvantaged, leading us to an intellectual labour automation crisis. Where jobs are replaced due to competitive forces and long-term health of humanity is thrown out the window. While I view this as an AI governance problem, I suspect that legal requirements might not be enough to control who has access to powerful AI, especially if the development process remains not contained. So I suspect that cryptographic schemes might be necessary to prevent that scenario.

Guaranteeing Safe AI via Boxing

Lastly, building on the idea proposed by Paul Christiano in 2010 to lock AI in a box and ask it to align itself

Lastly, I want to expand on the idea proposed by, which is to lock the AI in a box, and in some preconditioned way only let it out if proof is provided that the AI has aligned itself (i.e. created a copy of itself that it is aligned). If correctly implemented such a set-up would give us a guarantee that the AI is aligned, potentially mitigating the problems associated with the centralisation of power or misuse, as aligned superintelligence might be smart enough to refuse any actions that would lead to undesired outcomes to humans how don't have access to it). Asking an unaligned AGI to create an aligned version of itself and create proof that it is aligned is quite abstract. A more pragmatic version of this approach could be in the style of a framework outlined in the Towards Guaranteed Safe AI paper. The authors propose that by combining a *safety specification, a world model,* and *a verifier* we can quantifiably determine if a given AI system is safe.

We could box the unaligned AI and ask it to produce a world model, safety specification, and verifier (if we don't have these ourselves). The output from the unaligned AI is unlikely to pose much risk if it is safely contained, however, the examination of the given components by the AI would need to be thorough. Then we can ask the unaligned AI to

create a version of itself that satisfies the safety specification for a given world model and use the verifier to validate or invalidate the