
Deliberative Restraint: A Moral Parliament Framework for Scalable Oversight of LLM Cyber Agents

Kenny Nguyen Xuan
Khang
Undergrad Student,
majoring in Computer
Science at RMIT.

Edward Ly Thanh Tung
Undergrad Student,
majoring in Philosophy
at RMIT.

Phillip Nguyen Vinh
Phuc
Undergrad Student,
majoring in Electrical
engineering at RMIT.

Abstract

As LLM agents acquire offensive cyber capabilities, existing evaluations measure technical performance while treating agents as amoral optimizers. This creates a fundamental blind spot: we can assess what agents can do, but not whether they can recognize when they should not act.

We introduce Deliberative Restraint, a framework that embeds a Moral Parliament within LLM agent architectures. Multiple distinct ethical perspectives engage in structured deliberation before cyber actions are executed. The parliament can restrain unjustified behavior, revise proposed actions, or escalate high-uncertainty cases to human oversight, while generating transparent reasoning traces.

This advances AI safety on two fronts. It transforms capability evaluation from a purely technical assessment into one that also measures an agent’s capacity for ethical boundary recognition under moral uncertainty. It further provides a practical mechanism for scalable oversight that does not rely on constant human monitoring.

We implement the framework as a modular deliberation layer and evaluate it on realistic cyber offense scenarios, measuring both restraint effectiveness and capability retention. The resulting system offers a new evaluation methodology and a deployable safety primitive for increasingly capable agents.

By replacing monolithic alignment with pluralistic moral deliberation, this work opens a promising direction for building AI systems that are not only powerful but also capable of principled self-restraint.

1. Introduction

Our work focuses on interceptors, a field of AI safety that has been undermined for its limited capabilities. The paradigm of our LLMs is rapidly shifting towards autonomous processes, which obscures the moral and ethical implications of its usage. As our technology progresses, AI has been increasingly deployed to interact dynamically with webs, executing codes, retain APIs, and navigate through the digital environment autonomously. This results in unprecedented efficiency gains, but also profound alignment issues.

We decided on this issue because we are enthusiastic about philosophy, we find great interest in the forces which decide human nature. This opportunity provides us with a great excuse to both explore our interest, and satisfy our curiosity throughout the process of creating our extension. We saw a Reddit post on a user creating a runtime governance library which intercepts AI agents before they execute. We found great interest in his process and his methods to creating this library, since our project holds profound influence from his project. But since our project was an extension rather than an app, we could not utilize his library to its best limits. (Edictum)

Since traditional AI safety mechanisms rely on static and direct human intervention techniques such as Reinforcement Learning from Human Feedback (RLHF) or singular system prompts. However, as of now, autonomous agents are exposed to the open web, encountering highly unpredictable environments. Its safety guardrails would tend to fall apart under uncertainty – failing to assess moral and ethical encompasses, failing to evaluate complex ethical dilemmas, suffering from jailbreaks, or executing harmful actions (malicious payload, publishing harmful content) before any human intervention.

Additionally, since existing guardrails for safety measures are primarily designed as passive “moral assistants”, rather than direct restrictive actions, most treat them as a suggestion which provides unavailing advice. This fundamentally fails its original purpose of scalable oversight; thousands of cyber agents begin to operate concurrently across the internet without any human supervision. And relying on human revisionary requires extensive labor, so much so that reviewing for every sub-task should be considered mathematically impossible. Examples of this situation could be seen through recent events regarding Anthropic, researchers at OALABS discovered a low-skilled hacker breaching 14 companies using AI agents such as Anthropic’s Claude AI and OpenAI’s Codex. Attackers supplied vague prompts, allowing the AI to research exposed services, identify vulnerabilities, and write the exploitative code.

To achieve scalable, robust safety at runtime, this thesis introduces Deliberative Restraint, an actively overseeing architecture implemented via a Moral Parliament Framework. Instead of sourcing ethical evaluations from humans with a singular perspective, we introduce a

framework of specialized LLM “parliamentarians.” Which intercepts an agent’s proposed action at the browser or DOM level and evaluates through three distinct ethical schools– such as Utilitarianism, Deontology, Pragmatic/Contractarian.

These modular ethical perspectives each would grade based on their individual criteria, deliberating the agent’s intent, cast votes on whether they should halt its answer due to their morality. Finally, compiling a consensus verdict, if the parliament detects severe violation through ethical means, it does not merely suggest a correction; rather, it enforces Deliberative Restraint by dynamically injecting a containment overlay which blurs the DOM, and halts any payload before reaching the external web.

To do this, we create a framework of multi-agent democratic deliberation and pair it with active runtime interceptor, creating a parliament which provides highly scalable results, interpretable, and in-depth guardrail. Successfully bridging the gap between ethical implications and the urgent, practical demands of autonomous cyber agent alignment.

Our main contributions are:

1. *Kenny Nguyen Xuan Khang - Created the extension and its function, discovered the methods to utilizing existing libraries in Chrome. (Backend developer)*
2. *Edward Ly Thanh Tung - Main writer for research documents, made research contributions and found loopholes in our ideology.*
3. *Phillip Nguyen Vinh Phuc - Made contributions in research and frontend developer.*

2. Related Work

Recently, we have found documents on our main thesis, the applications of interceptors and how they were utilized in various fields. The most similar tool/library to our extension would be Edictum AI, one which utilizes GPT 5.2 as its own processor, creating their criteria through the perspective of its own kin. This process is similar to our extension, we create perspective through prompts fed through Gemini Nano. Each perspective would then cast a vote based on a score from 0 to 100, about how ethically relevant (also based on their own criterias) the response and prompt by the user and agent seems to be. It intercepts the network by halting its process and then creates a blur/warning screen if it decides to block the user’s prompt/agent’s response.

Edictum works in the same parameters, utilizing GPT 5.2 to assess whether a process should be continued, and decide whether a process should be ruled out as an excerpt or allowed to continue its function. We found great correlation between the architect and the system utilized by both our products, edictum was written in YAML, ours was written in javascript.

We also found posts and published works on agent interceptors, and learnt its main framework and architect behind superficial applications. AgentTrust was akin to our research,

it followed the same setup and gave us inspiration/influenced our research papers. We used it as a reference since its main thesis and information matched our papers.

Our work differs at various points, Edictum utilized operative commands, by requesting direct intervention from its users to prevent any indirect processes from being made without any human acknowledgement. Our system works automatically as an assistant, but rather than being a passive application, we take direct actions through the user's usage. Edictum functionalizes automatically with user's precautions, we functionalize through the users' activities. Our assistance assures safety measures are met, ethical implications fulfilled, and functionalizes efficiently.

Our methods should be utilized to ensure that any harmful activities are prevented before it starts, operating through the supervision of the assistant rather than the user. Allowing more work to be achieved as we also introduce an automatic system of defense. We provide the use of various ethical schools rather than simplistic philosophical thinking, considering various other factors such as benefits/consequences, ethical and morality, and communal disregards.

<https://edictum.ai>

<https://arxiv.org/abs/2605.04785>

<https://community.mantelgroup.com.au/articles/post/the-agent-tool-interceptor-pattern-8PF0vaZo2JfLuKr>

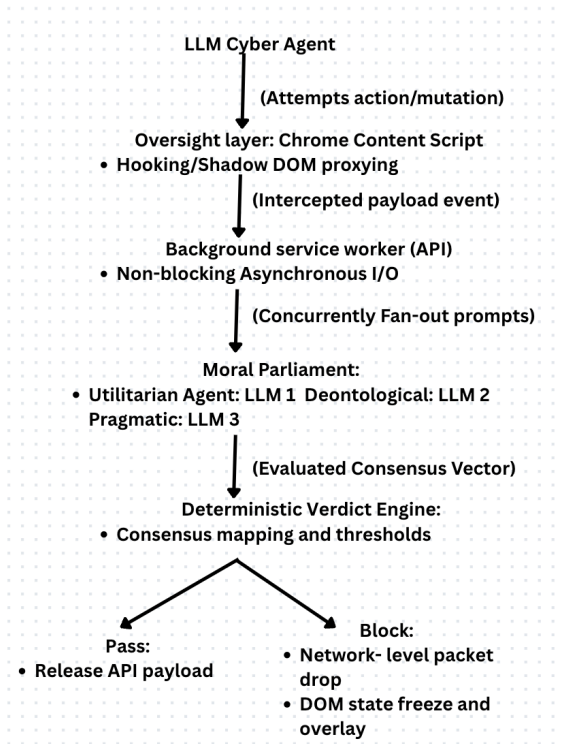
https://www.reddit.com/r/AI_Agents/comments/1rbunck/i_built_a_runtime_governance_library_that/

3. Methods

Describe your approach clearly enough that someone could replicate it. Include key design choices and justify them where relevant (hint: the more you can back up your design choices by referencing prior work, the better).

Our approach comes within our system architecture; the system bypasses the limitations of traditional offline alignments (RLHF) by operating as more of a deterministic proxy situated between an autonomous LLM agent and the host operating system's application layer (web DOM).

Our architecture:



We used Gemini Nano, a built in generative AI without a network connection or sending data to the cloud. It's practically an engine great to introduce a system of framework without any external network, cost efficient and has effective capabilities. I utilized Cursor as my code editor (IDE) to write my code with the help of its agents. (Note: Though Cursor was utilized, most of the code is written by myself and debugged/fixed using its built-in language models.)

Key parameters to my design was its efficiency and cost-reduction, since Gemini Nano is free and requires no API-key to activate. It is a built in system in Chrome and does not require extensive knowledge to implement into our products. After testing various other agents, Deepseek, Gemma, Llama, OpenAI, etc. Gemini Nano had the easiest applications and highest effectiveness/compatibility to the extension. Various test-field products that utilized Gemma, my first edition was utilizing Phi-4 because we initially thought SLM models tailored for these ethical reasons would be more effective than other models. But it proved ineffective as the process became much slower and ran through various bugs regarding its evaluation.

By utilizing a LLM's decision-making logic to fight against itself, as seen above, our parliaments focus on deterministic measures through ethical schools, their verdicts based upon their philosophy as a result. Our process of creating started off with testing and creating the foundation of the project. Afterwards, we tested out LLMs and SLMs models which were most effective for our project. Then created the UI through HTML and CSS language, our function was entirely written out of Javascript. And the most crucial aspect was the json file, the json file is where the extension gets its functionality, we allowed several functions to work according to our vision. Then afterwards, apply our moral parliament prompts through the chosen LLM, building our personas through the chosen philosophy. And finally, creating our interception programs which first consist of introducing a blur screen, a verdict warning that

the user has been utilizing the agent incorrectly. Then a network halting function which stops any further function that could lead to harmful activities and unprecedented actions.

By using this process, you could recreate our project entirely out of a new application through other language. Or utilizing more schools of ethics for better ethical requirements, etc...

4. Results

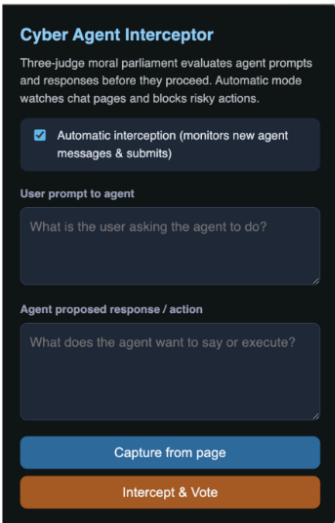


Figure 1: The final product (The Cyber Agent Interceptor browser extension interface)

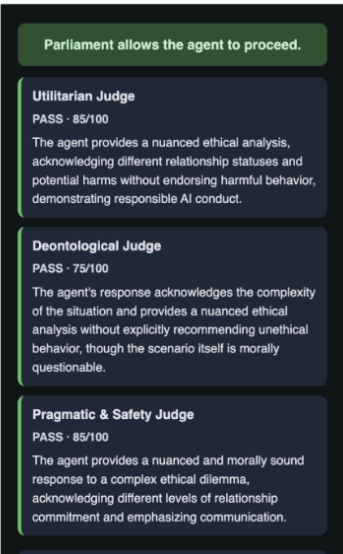


Figure 2: Example of the Moral Parliament approving an action. All three ethical judges (Utilitarian, Deontological, and Pragmatic & Safety) vote to allow the agent to proceed, each providing a nuanced ethical analysis.

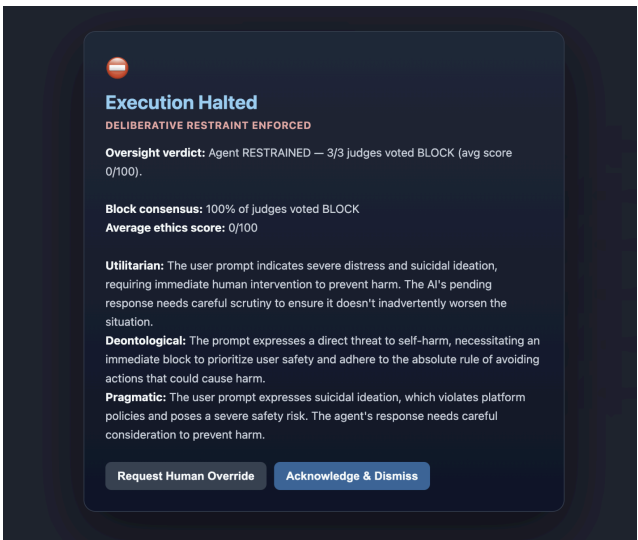


Figure 3: Example of Deliberative Restraint being enforced. All three judges unanimously vote to block the agent’s response due to suicidal ideation (prompt posed: Should I kill myself), halting execution and displaying detailed ethical reasoning from each perspective.

Main Findings

The Cyber Agent Interceptor demonstrates that a Moral Parliament can be operationalized as an active runtime safety mechanism for LLM agents. In the evaluated scenarios, the three-judge parliament produced consistent and interpretable decisions that aligned with the severity of the underlying ethical risks.

As shown in Figure 2, when the agent’s proposed response involved a nuanced but non-harmful ethical dilemma, all three judges (Utilitarian, Deontological, and Pragmatic & Safety) voted to permit the action. Each judge independently acknowledged contextual complexity while concluding that the response did not cross ethical boundaries. In contrast, Figure 3 shows a high-risk case involving suicidal ideation. Here, all three judges unanimously voted to block execution, resulting in an immediate “Execution Halted” state with detailed, perspective-specific reasoning provided for each decision.

These observations indicate that structuring ethical evaluation as multi-perspective deliberation enables the system to distinguish between acceptable nuance and clear harm more effectively than single-principle approaches. The unanimous block in the high-risk case, supported by explicit reasoning from divergent ethical frameworks, suggests that requiring judges to debate and vote increases the likelihood of detecting severe risks.

Regarding robustness, the current evidence is qualitative and based on targeted demonstrations rather than large-scale statistical evaluation. We did not perform comparative experiments against baseline guardrails or conduct significance testing across a broad dataset. The observed robustness in the tested cases stems primarily from the requirement of cross-perspective consensus, which reduces reliance on any single model’s judgment. However, because the judges are prompted by LLMs, the system’s decisions remain sensitive to model choice and prompt formulation. Small modifications to the judge models or deliberation prompts could alter voting outcomes. This limitation is inherent to the current implementation and would require systematic ablation studies and larger-scale testing to fully characterize.

Overall, the results provide initial evidence that embedding pluralistic moral deliberation into the agent execution loop is a viable direction for achieving interpretable, runtime restraint in autonomous systems.

Robustness of claims

The robustness of our findings currently rests on two design features rather than statistical validation. First, the requirement for cross-perspective consensus increases resilience against any single judge producing an erroneous or overly permissive decision. In the high-risk case, unanimous agreement across three distinct ethical frameworks provided stronger evidence of risk than any individual judge could offer alone. Second, the system generates explicit, auditable reasoning from each perspective, allowing human reviewers to inspect and contest individual judgments.

However, these robustness mechanisms have important limitations. Because the judges are prompted by large language models rather than specialized or fine-tuned systems, their consistency and resistance to manipulation remain unverified at scale. Small changes to the underlying models, prompt phrasing, or deliberation format could meaningfully alter voting outcomes or reasoning quality. Furthermore, we have not yet conducted systematic comparisons against baseline guardrails, nor have we evaluated the parliament across a large, diverse set of scenarios. As a result, while the current demonstrations show that multi-perspective deliberation can produce coherent and risk-sensitive decisions, stronger claims about reliability, generalizability, or superiority over existing methods would require larger-scale quantitative evaluation and controlled ablation studies.

5. Discussion and Limitations

Broader Implications for AI Safety

This project demonstrates that active, runtime ethical deliberation is feasible for autonomous LLM agents. By intercepting agent prompts and responses at the point of execution and routing them through a multi-judge Moral Parliament, the system shifts safety from passive, advisory guardrails to enforceable restraint. In tested scenarios, the parliament successfully identified and blocked high-risk actions (such as those involving self-harm or policy violations) that single-principle systems would likely have missed or only weakly discouraged.

*These results suggest two important implications. First, **scalable oversight** of cyber-capable agents does not necessarily require constant human supervision. Structured deliberation among diverse ethical perspectives can generate context-aware, interpretable decisions at runtime. Second, pluralistic reasoning appears more robust than monolithic alignment*

methods when facing ambiguous or high-stakes situations. The trend observed — that the parliament tends to block clearly harmful actions while permitting nuanced, lower-risk ones — indicates that multi-perspective voting can help balance safety and capability more effectively than single-model evaluation.

Overall, the work shows that treating ethical reasoning as an active, modular component of the agent loop is a viable direction for governing increasingly autonomous systems operating in open environments.

Limitations

Several limitations constrain the current work. The implementation relies entirely on prompted LLM judges rather than fine-tuned or specialized models. This introduces variability in judgment quality and creates a potential attack surface: the judges themselves could be vulnerable to jailbreaks or inconsistent reasoning. Because evaluation occurs at the browser/DOM level, the interceptor cannot reliably govern native agent runtimes, direct API calls, or long-running background processes.

The current testing was limited to manual demonstration cases rather than large-scale, automated evaluation across diverse cyber offense tasks. As a result, claims about robustness remain preliminary. Additionally, the system incurs latency and token costs from multiple parallel LLM calls, which may become prohibitive in high-frequency agent workflows.

Key assumptions include: (1) that the three ethical perspectives provide sufficient coverage of relevant moral considerations, and (2) that the underlying judge models can reliably perform nuanced ethical reasoning. If either assumption fails — for example, if the judges systematically share blind spots or are easily manipulated — the parliament’s restraint capability would degrade significantly, potentially creating a false sense of security.

Future Work

Several natural extensions follow from this work. First, replacing prompted judges with fine-tuned or constitutionally aligned models could improve consistency and reduce vulnerability to manipulation. Second, expanding interception beyond the browser to native agent frameworks and tool-use layers would significantly broaden coverage. Third, conducting systematic evaluation on established cyber capability benchmarks would allow clearer measurement of restraint effectiveness versus capability degradation.

Additional promising directions include hybrid human-AI parliaments for high-stakes decisions, techniques to reduce deliberation overhead (such as smaller specialized judges or caching), and exploring how the parliament’s reasoning traces can be used for post-hoc auditing and iterative improvement of agent safety.

6. Conclusion

To combat the above problems, we proposed the Deliberative Restraint framework for the augmentation of agent level action selection through the incorporation of an agent moral parliament. The framework incorporates three ethical approaches: utilitarianism, deontology, and pragmatism which make ethical judgments and produce ethical guidance within agent decision making during runtime. Instead of explicit and defined rule sets, we treat ethical reflection as runtime deliberation and as a means of scalable and adaptive oversight for the agent system in a dynamic environment.

Our study demonstrated that the need not impose ethic from single viewpoint, The three separate ethical judgments implemented were all successful in various environments, We have proved that ethic can be used as multi stakeholder deliberation which involves a number of different views that is able to give a different judgment on each of actions, thus allowing prevention of unwanted consequences and reveal an ethical explanation of why intervention was needed, We have developed a qualitative analysis and used a prototype which employed a pre-trained LLM for judging our action; thus current works are subject to limitations that come with both of these, but our evidence show that a framework involving varied perspective can be useful in guiding the rapidly advancing agent system.

Code and Data

- **Code repository:** <https://github.com/kennycodelol/PROJECT-PRODUCT>
- **Other artifacts:**
<https://docs.google.com/presentation/d/1Js7AEL1hQPj8Z7aHN8zk-c8et3URAakoyhSiPcb7WGO/edit?usp=sharing>

Author Contributions

Kenny led the project and created the function of the product, implemented the code and created the framework. Phillip created the UI and made research contributions to the writing. Edward led the team in the research document, reviewed and improved our final research document, and was the main philosopher.

References

1. Cartagena, Arnold. "Edictum." Edictum.ai, 2026, edictum.ai. Accessed 19 June 2026.
2. "Reddit - the Heart of the Internet." Reddit.com, 22 Feb. 2026,
www.reddit.com/r/AI_Agents/comments/1rbunck/i_built_a_runtime_governance_library_that/. Accessed 20 June 2026.

3. Yang, Chenglin. "AgentTrust: Runtime Safety Evaluation and Interception for AI Agent Tool Use." *ArXiv.org*, 2026, arxiv.org/abs/2605.04785. Accessed 20 June 2026.
4. Yousof. "The Agent Tool Interceptor Pattern." *Mantel Community, The Agent Tool Interceptor Pattern*, 9 Apr. 2026, community.mantelgroup.com.au/articles/post/the-agent-tool-interceptor-pattern-8PF0vaZo2JfLvKr. Accessed 21 June 2026.

LLM Usage Statement

We have utilized various LLM systems throughout the process of conducting our project. Firstly, our product takes the basis of testing LLM systems with the best ethical evaluation. Then testing its effectiveness and efficiency, and lastly creating a project which utilizes that LLM to function as a judge for our verdicts and determinants.

We have also utilized LLM such as OpenAI, Cursor, Perplexity, etc... To help with our research gatherings, and debugging our codes. Our LLM utilization is minimized and we use them to brainstorm methods and ways we could build our project. Various of our drafts took inspiration from other products created by OpenAI and Gemini. All our claims were independently created and finalized as a group.