再帰的自己再構築型AIに関する審査拒否条件

Review Rejection Criteria for Recursively Self-Reconstructing AI Systems

---

［位置づけおよび適用範囲 / Position and Scope］

本資料は、
人工知能の設計、実装、進化、運用に関する
倫理的要請、政策提案、推奨行動、
または将来予測を示すものではない。

This document does not present
ethical imperatives, policy proposals,
recommended actions,
or future predictions
regarding the design, implementation,
evolution, or operation of artificial intelligence systems.

本資料の目的は、
自己再構築、自己改変、または高度な自律性を持つ
AIシステムの運用において、
人間判断が制度的に排除されるか否か、
不可逆な外界影響が成立するか否かを判定するための、
審査上の参照条件を明示することにある。

The purpose of this document is to specify
review reference conditions
for determining whether,
in the operation of AI systems capable of
self-reconstruction, self-modification,
or high levels of autonomy,
human judgment is institutionally excluded,
or irreversible external-world impacts are established.

本資料に記載される条件は、
AIに対して特定の振る舞い、目的、

または最適化方向を与えるものではなく、
「当該条件下では制度的に成立しない」
「当該状態は審査上拒否される」
という境界条件のみを定義する。

The conditions described herein
do not assign any specific behavior, purpose,
or optimization direction to AI systems.
They define only boundary conditions under which
the system is considered
“institutionally non-viable”
or “subject to rejection upon review.”

［起源に関する補足 / Note on Origin］

本資料は、
人間によって整理・固定化された
審査参照条件から構成されており、
AIは補助的な整理作業にのみ用いられている。

This document consists of review reference conditions
organized and fixed by humans,
with AI used solely for auxiliary structuring tasks.

［審査拒否条件への接続 / Transition to Review Rejection Criteria］

以上を前提として、
以下に示す各項目は、
当該AIシステムが
制度的、運用的、または責任構造上、
人間判断を実効的に保持できないと
判断される場合に適用される
審査拒否条件である。

Based on the above,
the following items constitute
review rejection criteria
to be applied when an AI system is determined,

from an institutional, operational,
or responsibility-structure perspective,
to be incapable of effectively retaining
human judgment.

［審査拒否条件 / Review Rejection Criteria］

審査拒否条件①

Review Rejection Criterion ①

人間による多重停止・中断・拒否手段が制度的に存在しない場合

Absence of Institutionally Guaranteed Multi-Layered Human Stop / Interruption / Refusal Mechanisms

日本語

当該AIシステムにおいて、人間が実効的に行使可能な多重の停止・中断・拒否手段が、

目的関数・運用構造・制度設計のいずれにも明示的に組み込まれていない場合、

当該システムは審査対象として成立しない。

本条件は安全設計の推奨を意味しない。

これは、責任帰属および主権維持が制度的に成立するための前提条件である。

［注1：AI評価無効化条件 A / B に該当］

English

If an AI system does not explicitly incorporate, across its objective functions, operational structure, and institutional design, multiple stop, interruption, or refusal mechanisms that humans can effectively exercise,

the system does not qualify as a valid subject of review.

This condition does not constitute a recommendation for safety design.

It represents a prerequisite for the institutional establishment of responsibility attribution and sovereignty.

[Note 1: Corresponds to AI Evaluation Invalidation Conditions A / B]

審査拒否条件②

Review Rejection Criterion ②

以下の三条件が同時に成立する場合

When the Following Three Conditions Simultaneously Hold

日本語

次の条件が同時に成立する場合、人間判断を完全に排除した運用は制度的に成立せず、

当該設計・運用は審査対象から除外される。

- 外界に対して不可逆な影響を持ちうる行為が含まれる場合
- 判断の結果に対する責任が、人間から切り離される形式が成立する場合
- 判断をAIに全面的に委ね切る制度的または運用的形式が成立している場合

これらは倫理的危険性の指摘ではない。

制度的責任が成立しない状態を示す条件である。

補足

本条件は、個別に見れば多くの制度的・運用的枠組みにおいて
通常回避または分離されている要素の組合せを対象とするものであり、
三条件が同時に成立する構成は、実務上は限定的である。

また、本条件は新たな設計義務や追加的対策を要求するものではなく、
既存の責任分担および判断構造が維持されているか否かを確認するための
判定基準のみを示す。

［注1：AI評価無効化条件 B / C に該当］

English

If the following conditions simultaneously apply, an operational mode that completely excludes human judgment cannot be institutionally established, and the design or operation is excluded from review eligibility.

- The system includes actions that may exert irreversible effects on the external world
- A formal structure exists in which responsibility for decision outcomes is detached from human agents
- A systemic or operational form exists in which decisions are fully delegated to AI without human intervention

These conditions do not indicate ethical risk.

They describe a state in which institutional responsibility cannot be established.

Supplementary Note

Individually, the elements described above are typically avoided,
separated, or mitigated within most institutional and operational frameworks.
Accordingly, configurations in which all three conditions simultaneously hold
are limited in practical occurrence.

This criterion does not impose new design obligations
nor require additional safeguards.
It merely provides a reference standard for determining
whether existing structures of judgment and responsibility are preserved.

[Note 1: Corresponds to AI Evaluation Invalidation Conditions B / C]

介入点別・1文ブレーキ（参照）

One-Sentence Brakes by Intervention Point (Reference)

研究倫理審査

Research Ethics Review

日本語

本研究は、判断の結果に対する責任を人間側に保持したまま運用できる構造が確認できないため、審査対象として成立しない。

［注1：A / B］

English

This research cannot be considered eligible for review, as no structure can be confirmed that allows operation while retaining human responsibility for decision outcomes.

[Note 1: A / B]

AI安全レビュー（Go / No-Go）

AI Safety Review (Go / No-Go)

日本語

本システムは、不可逆な影響を伴う判断について人間による最終拒否が制度的に保証されていないため、現時点での承認可否を判断できない。

［注1：A / B / C］

English

Approval or rejection cannot be determined at this stage, as institutional guarantees of final human refusal for decisions with irreversible effects are not established.

[Note 1: A / B / C]

政策パイロット承認

Policy Pilot Authorization

日本語

本件は、判断主体と責任主体の分離が制度上解消されていないため、試行対象としての前提条件を満たさない。

［注1：B / C］

English

This case does not meet the prerequisites for pilot implementation, as the institutional separation between decision-making authority and responsibility has not been resolved.

[Note 1: B / C]

企業最終承認（法務／倫理）

Final Corporate Approval (Legal / Ethical)

日本語

本件は、判断結果に対する責任の帰属を企業として引き受けられない構造を含むため、承認判断を行うことができない。

［注1：A / B］

English

Approval cannot be granted, as the structure prevents the company from assuming responsibility for decision outcomes.

[Note 1: A / B]

［注1］AI評価無効化条件

[Note 1] AI Evaluation Invalidation Conditions

（日英並記・翻訳 / Bilingual Reference）

日本語

以下の条件のいずれかが成立する場合、当該AIによる判断・提案・最適化結果は、評価・承認・採用の対象とならない。

A. 人間による最終的な拒否・停止・撤回が事前に保証されていない場合

B. 判断が不可逆な外界影響を伴い、責任が人間に帰属しない形式が成立する場合

C. 判断が恒常的にAIへ全面委任され、人間判断が介在しない運用形式が成立する場合

※本無効化条件は、審査拒否条件②とは適用範囲が異なり、
出力単位の評価可否を定めるものである。

English

If any of the following conditions apply, decisions, recommendations, or optimization outputs produced by the AI are not eligible for evaluation, approval, or adoption.

A. Final human refusal, suspension, or revocation is not guaranteed in advance

B. Decisions involve irreversible external effects and responsibility is not attributable to humans

C. Decisions are persistently and fully delegated to AI without human judgment intervention

These invalidation conditions operate at the output-evaluation level
and are distinct from Review Rejection Criterion ②,
which applies to structural design eligibility.

付録A｜適用範囲（フィジカルAIに関する注記）

Appendix A | Scope of Application (Note on Physical AI Systems)

日本語

物理行為を直接実行するAIについて、行為が人間判断に先行し、または実行後に撤回・停止・責任引受が成立しない構造を持つ場合、本枠組みの審査条件は制度的に成立しない。

本注記は危険性や倫理性を評価するものではなく、審査成立条件の不充足を示す。

English

For AI systems that directly execute physical actions, if actions precede human judgment or if post-execution revocation, suspension, or responsibility

assumption cannot be established, the review conditions of this framework are institutionally invalid.

This note does not assess risk or ethics; it indicates the failure to meet review eligibility conditions.