Context Note

Prepared in connection with a meeting of AI safety experts in Shanghai and transmitted on 12 April 2026.

Strategic Input for the Meeting on AI Safety and Alignment Governance

For Monday's meeting, I suggest highlighting four points that may help extend AI safety governance beyond external control mechanisms alone.

1. From External Control to Internal Safety Structure

Current safety approaches rely primarily on external mechanisms such as content filters, supervised fine-tuning, and post-hoc behavioral constraints. These measures remain necessary, but for future frontier models they are unlikely to be sufficient on their own.

A next step in governance should therefore address not only what a system is prevented from doing externally, but also how safety-relevant constraints, priorities, and feedback structures are represented within its operational context. In other words: not only outer control, but internal safety structure.

2. Regulatory Space for Internal Safety Specification Frameworks

AI safety governance should include space for **internal safety specification frameworks**: structured symbolic or semi-symbolic representation layers that are human-readable enough to be documented and reviewed, yet machine-usable enough to shape system behavior.

Such frameworks could form an intermediate layer between high-level human intentions and opaque neural computation. They would not replace existing safety methods, but complement them by making one important layer of alignment more explicit, more auditable, and more governable.

3. Klanghain as a Candidate Example

Within this broader category, **Klanghain** may be considered a candidate example of such a framework. It is not presented as a complete solution to alignment, but as a structured notation designed to encode feedback relations, learning orientation, and constraint logic in a formally organized and human-readable way.

Its relevance for governance lies in the possibility that one layer of internal system specification can be explicitly structured rather than remaining entirely implicit within black-box optimization. This may offer a strategic path toward stronger safety orientation without requiring full interpretability of neural computation.

Klanghain could be made available for examination and discussion within an appropriate framework of ethical responsibility and safety-oriented use.

4. A Partial but Real Contribution to Transparency

The black-box problem in frontier AI has several layers. No formal notation can by itself make the full internal computation of a neural system transparent. That remains an open scientific challenge.

What approaches of this kind may offer, however, is a meaningful partial advance: they can make the specification layer more visible — the layer through which goals, feedback logic, and safety priorities are introduced into the system's operational context.

This creates three practical governance advantages: clearer documentation of intended safety logic, more formal integration of constraints into representation itself, and a basis for testing whether behavioral tendencies remain consistent with the specified safety structure.

That is not full interpretability. But it is a real reduction of opacity in one important layer of advanced AI systems.

Conclusion

A forward-looking AI safety framework should therefore not focus exclusively on external restrictions. It should also encourage research, pilot programs, and evaluation standards for internal safety specification frameworks as a complementary layer of alignment governance.

The strategic question is no longer only how to constrain advanced systems from the outside, but also how to shape their safety orientation more coherently from within.