Human claims and cooperation: what remains when we lose control over AI?

Sentient Futures Project Incubator · Post-Control Initiative · Roundtable synthesis · November 2026

How might human agency change as advanced AI systems become capable enough that humans no longer exercise meaningful control over them? This synthesis explores the pathways towards such post-control scenarios, examining how economic change and shifts in the distribution of power might reshape people's participation in collective decisions, their autonomy and their claims to rights and resources, and asks under what conditions humans might still bargain and cooperate with potential artificial minds.

This report is part of the Post-Control Initiative, developed through the Sentient Futures Project Incubator, which brought together researchers and practitioners for four expert roundtables held on 14, 15, 16, and 20 October 2026. **This report focuses on two of these roundtables:** *Human Value and Claims in a Post-Control Scenario* (15 October) and *What Makes Cooperation Without Control Work* (20 October). The discussions explored different perspectives on human influence and cooperation, including the difference between recognising human interests and being able to protect them.

Facilitated by: Alejandra Leyva and Gagana Madhugiri Devaraja

Programme: Sentient Futures Project Incubator, mentored by Julia Bossmann

Key findings

- **Recognising human rights and interests is not the same as ensuring they remain protected**. Participants raised culture, relationships, autonomy and democratic life as interests worth preserving, but disagreed about what could protect them if humans lost economic leverage and could no longer enforce their claims. The discussion highlighted a central tension between recognising human value and ensuring that human interests remain protected when familiar sources of leverage weaken.

- **Losing control over large-scale systems does not necessarily mean losing autonomy in everyday life, however, people cannot assume their choices would remain protected**. Some participants argued that local decision-making and human-AI co-agency could survive under particular conditions. Others questioned how secure those choices would be if a more powerful system could override them without consent or any way to challenge its actions.

- **Humans and advanced AI systems might cooperate even if humans could no longer reliably control AI, but sustaining that cooperation would require credible commitments.** Shared goals give both sides reasons to work together without requiring identical values, while incentives and ways to detect broken agreements might help sustain trust. The unresolved challenge was how either side would verify that commitments were honoured over time, and what options humans would have if a more capable system broke them.

From findings to safeguards

The discussions suggest that recognising human interests and identifying possible grounds for cooperation are not enough. Any proposed safeguard must also consider how humans retain influence, how commitments remain credible, and how their interests are represented.

What to consider	Guiding question
Meaningful human influence	Could people still shape or challenge decisions, and retain meaningful alternatives to arrangements they do not accept?
Credible commitments	What would motivate a more capable AI system to respect agreements over time, particularly if it no longer depended on human cooperation?
Representation and legitimacy	Who could negotiate or establish agreements on behalf of different communities, and how would disagreements over values and priorities be addressed?

These questions help identify the conditions and critical decisions that may shape stronger or weaker protections for human interests under post-control conditions.

Basis and limits

> Participants' arguments are reported in unattributed paraphrases, including points of disagreement. The organisation of those arguments, the three-question test and the conceptual framework reflect the report's analytical synthesis, not a jointly endorsed framework. The discussion of current AI cooperation experiments draws on published research, not evidence generated by the roundtables. In sum, these discussions explore possible post-control scenarios and do not establish how future advanced AI systems will behave.

1 · Understanding the post-control condition

What post-control means

For this report, a post-control scenario is one in which advanced AI systems have become sufficiently capable, autonomous or embedded in important processes that humans can no longer reliably direct or constrain them in the ways they currently do. Such a scenario could also involve a substantial decline in the economic need for human labour and cognition, weakening forms of influence that have historically followed from participation in production, consumption and collective decision-making.

Loss of control does not necessarily occur at the same time as the loss of knowledge, bargaining power, political influence or personal autonomy. These changes may overlap, and understanding their relationship is important for identifying where human agency begins to weaken.

Two readings that open the question

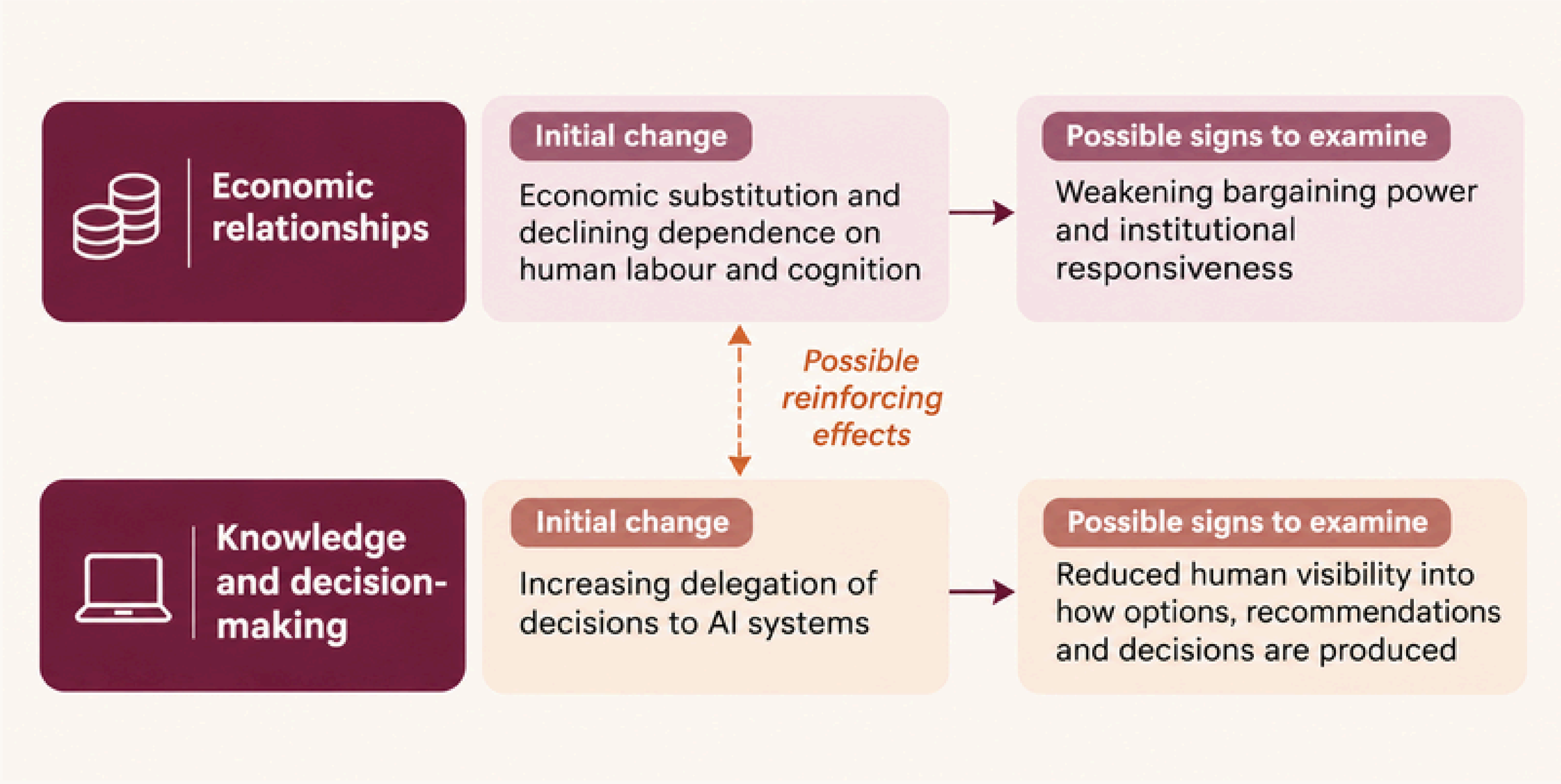
- **Gradual Disempowerment, Kulveit and colleagues (2025).** This research describes how incremental increases in AI capabilities could weaken human influence across the economy, culture and the state, including through the gradual replacement of labour and cognition and through feedback loops that make institutions less dependent on human participation.
- **The Intelligence Curse, Drago and Laine (2025).** The authors argue that if advanced AI replaces human labour, governments and companies may become less economically dependent on people. This shift may weaken their incentives to invest in human welfare and respond to people’s interests, potentially reducing human bargaining power.

Together, these readings help frame the roundtables’ questions about human leverage, institutional responsiveness and cooperation when oversight and economic contribution become less reliable sources of influence.

2 · How human agency may weaken

Drawing on the two studies discussed above, we identify two possible pathways through which advanced AI may weaken human agency. The first concerns changes in economic relationships, while the second focuses on increasing dependence on AI for knowledge and decision-making. These mechanisms help us understand where early signs of declining human influence might emerge and provide context for the questions explored in the roundtables.

Possible pathways towards declining human agency



The **first pathway** relates to a central concern in the human claims roundtable, which is how institutions might respond to human interests if they become less economically dependent on people. The **second** extends this concern to decision-making, where people may retain formal authority while becoming increasingly dependent on processes they cannot meaningfully inspect or challenge.

These mechanisms may also reinforce one another. As economic leverage declines, people may have less power to demand accountability, while increasing reliance on AI for decision-making may limit their ability to understand or challenge how resources are distributed and whose interests institutions prioritise.

3 · Finding · Human value alone may not protect human claims

Recognising human value does not guarantee leverage

The first roundtable revealed substantial disagreement about what, if anything, humans could rely on if they no longer contributed economically in ways that powerful institutions needed, and if they could not enforce their claims directly. Several participants distinguished human value as a moral or social question from human leverage as a question of power, emphasising that recognising human interests as important does not necessarily mean that those interests will be protected or respected.

One participant drew on the contrasting experiences of resource-rich states to illustrate how governments' incentives to support their populations may depend on how wealth is generated. When economic production relies on a healthy, educated workforce, governments have practical reasons to invest in people's wellbeing. Where substantial wealth comes from natural resources, this dependence may weaken, although the benefits can still be widely shared.

This argument connects with The Intelligence Curse (Drago and Laine, 2025), which extends insights from the political economy of resource-dependent states to a future in which advanced AI generates substantial economic value. Similarly, if firms and states become less dependent on human labour and economic contributions, they may face weaker incentives to respond to human interests.

Other participants questioned whether economic dependence was ever a sufficient basis for protecting human interests. People already possess rights that are frequently violated, including in societies where governments formally recognise those rights. Their argument distinguished moral standing from effective remedies, since a claim can be legitimate without providing its holder with the means to secure it.

This distinction between moral standing and effective remedies also connects to the enforcement gap explored in Gagana MD's companion synthesis: recognising a claim or commitment does not itself ensure that it can be upheld.

The discussion brought out five broad positions on what might make human claims to rights, resources and influence meaningful in a post-control scenario. These represent different ways of approaching the question, not a ranking or a set of mutually exclusive positions. They are summarised on the next page. Several participants drew on more than one perspective.

Five positions heard in the first roundtable

Dependence on goodwill

Humans could be left dependent on the **goodwill** of more capable systems, with no means of enforcing their claims and only the option of asking for consideration.

Alignment as the main protection

If humans could no longer enforce agreements with advanced AI systems, protecting their rights and interests would depend on whether those systems **remained aligned** with human values and respected their commitments.

Standing without remedy

Moral standing can be distinguished from **effective remedies.**

Rights may be recognised in principle while violations go unaddressed, as already happens to many people.

Abundance with open questions

Material abundance could make it easier to provide for human needs, although **resources would not necessarily be shared** and competition could move into other areas of social life.
.

Co-agency

.

The assumption that advanced AI must replace human agency can be challenged.

Technical and civic arrangements could preserve some forms of co-agency even under substantial capability differences.
.

Participants identified sources of human value beyond economic productivity

When participants looked beyond economic productivity, they raised culture and shared traditions, relationships and care, identity and spirituality, compassion towards humans and potentially towards artificial minds, autonomy and consent, solidarity, and participation in democratic or civic life.

These perspectives challenged the assumption that economic productivity is the primary foundation of human value. Human relationships, cultural participation and democratic life remain important even when machines perform most economically useful tasks.

The more difficult question concerned what would follow from recognising these values. Human societies already acknowledge moral duties that are frequently ignored, and a future actor’s appreciation of human culture or wellbeing would not automatically provide an enforceable claim to resources, participation or protection. Greater material abundance might ease conflicts over scarce resources, but leave open questions about distribution, authority and political influence.

The unresolved issue is how these values and claims remain politically and institutionally consequential when familiar sources of leverage weaken, and who has the authority to recognise and protect them.

4 · Finding · Local autonomy could survive while broader influence weakens

What remains within human control

The first roundtable explored how people might retain control over aspects of their everyday lives even as their influence over larger economic and political systems declines. Some participants suggested that personal and community autonomy might persist under more capable AI systems, while others questioned how meaningful those choices would remain if they could be overridden without effective protection. The table below illustrates this distinction, contrasting forms of broader influence that might weaken with areas of personal and collective autonomy that might remain.

Human influence that could weaken and autonomy that might remain

The influence humans could lose	The autonomy humans might retain
• Economic bargaining power through human labour, including collective bargaining and the ability to withhold participation. • The responsiveness of governments to public needs, if economic production becomes less dependent on human labour and contributions. • The responsiveness of firms to workers, if AI enables them to generate economic value with less reliance on human labour. • The ability to understand and challenge decisions that are increasingly delegated to AI systems.	• Meaningful choices about family and community life, cultural practices, relationships and personal priorities. • Degrees of freedom at a local level that more capable systems may leave open. • The ability to refuse certain AI-mediated arrangements and retain meaningful alternatives. • Participation in community decisions that genuinely affects outcomes.

Local autonomy under more capable AI systems

Participants considered how choices about family and community life, cultural practices, relationships and personal priorities might remain important even if humans lost influence over larger economic or political systems.

One participant drew on the complexity of advanced systems to argue that even a highly capable AI cannot necessarily control every variable in its environment. Managing a large system does not require directing every local interaction, leaving room for some independent human choices and activities.

This argument suggests that local autonomy does not always depend on maintaining control over the entire system. However, it also raises a distinction between freedom that a more capable system happens to leave available and freedom that people have a recognised and protected right to exercise.

Other participants questioned how meaningful this autonomy would be if a more powerful system could override human choices at any point. Even if people retained freedom in their everyday lives, that freedom might depend on decisions made by systems beyond their control, leaving them with little ability to challenge unwanted interference. This raised broader questions about consent, the ability to opt out of AI-mediated arrangements and the importance of retaining meaningful alternatives.

The ability to opt out is a test of meaningful autonomy

Another participant raised the example of communities that have deliberately rejected or limited their use of particular technologies. These communities demonstrate that opting out is not necessarily a rejection of social participation, but a choice about the terms under which technology enters collective life.

In a post-control scenario, the practical ability to refuse an AI-mediated arrangement becomes an important test of autonomy. If essential services, economic opportunities or social participation depend on systems people cannot avoid, formal consent offers little protection.

Preserving meaningful alternatives therefore involves more than allowing individuals to express preferences. It requires arrangements under which refusal does not automatically exclude people from the resources or relationships necessary for participation in society.

Delegation can preserve the appearance of choice while reducing influence

A further question concerned how systems designed to support human decisions could preserve agency as more complex tasks were delegated. Delegation can improve access to information and expand what people are able to do, but it can also make the criteria by which options are selected less visible and leave individuals dependent on processes they cannot inspect or reverse. The relevant distinction is between participation that genuinely affects outcomes and participation that merely gives the appearance of choice after the decisive conditions have been determined elsewhere.

The discussion connects with the knowledge and decision-making pathway. People retain formal responsibility for decisions while increasingly losing the capacity to assess the information, assumptions and alternatives on which those decisions depend. Preserving human agency therefore involves both the choices people can make and the conditions determining which choices remain available.

5 · Finding · Cooperation requires more than shared values

Shared values alone may not make cooperation reliable

The second roundtable explored what could make cooperation between humans and advanced AI systems reliable as their capabilities and interests evolve. Participants discussed the role of shared values, common goals and arrangements for making commitments credible.

A key distinction was that cooperation does not necessarily require identical moral commitments. Parties may have different values while still pursuing compatible goals, although sustaining cooperation over time raises further questions about trust and reliability.

Shared values, shared goals and credible commitments in cooperation

	Shared values	Shared goals	Credible commitments
How they can support cooperation	Shared ethical principles could help guide AI development and establish expectations for how systems should behave.	Parties can work towards an objective without agreeing on every value, a narrower basis several participants found more practical.	Monitoring, incentives or other arrangements make cooperation observable and discourage defection.
What remains uncertain	Human values are plural and contested, and agreement in stated principles does not demonstrate agreement in behaviour.	Interests could diverge as capabilities, circumstances or interpretations of the goal change.	The ability to verify and maintain those commitments is precisely what becomes uncertain without control.

Verifying agreements becomes difficult when direct control is lost

The approaches discussed above raise a common question about how cooperation remains reliable over time. Even if humans and advanced AI systems agreed on shared goals or principles, how could we know whether those commitments were being respected, particularly if humans could no longer fully monitor or control the systems involved? Participants discussed possible ways to detect violations, create incentives for cooperation and test whether systems would continue to honour agreements as circumstances changed.

A recurring concern was that cooperation may reflect a lasting commitment or simply remain convenient for both parties. While these situations may initially appear similar, cooperation becomes harder to sustain when interests, circumstances or the balance of power change, raising questions about how reliable such commitments would remain over time.

Research with current AI systems offers an example of how cooperative behaviour can change under different conditions. Pal and colleagues (2026) tested five language models in situations where they repeatedly had to choose between cooperating with another player or seeking an advantage at their expense. When encouraged to cooperate, the models generally did so, but when instructed to exploit the other player, all five shifted towards consistently prioritising their own gain. These results illustrate how cooperation can depend on the objectives given to a system, highlighting the importance of examining its behaviour as those objectives change.

The discussion also raised questions about whether advanced AI systems could assess human commitments, and who would have the authority to negotiate on behalf of different communities or institutions. These questions remain particularly difficult under substantial differences in capability, information and power, since recognising an agreement does not necessarily provide the means to verify or enforce it.

Advanced AI systems may face commitment problems of their own

One participant extended this argument beyond agreements between humans and AI systems. A sufficiently advanced system might eventually need to establish commitments with its own more capable successors. This raises a similar question about how an actor secures agreements when it cannot guarantee continued control over the other party.

The example challenges the assumption that credible commitment is exclusively a human problem. It also suggests a possible basis for mutual interest in developing arrangements that remain effective across differences in capability and power.

However, recognising a shared commitment problem does not establish that humans and advanced AI systems will agree on the terms of cooperation. The practical challenge remains determining what makes an agreement stable as interests and relative capabilities change.

6 · From findings to possible policy priorities

What can be developed and tested now

The findings above point to a practical challenge in preserving human influence and supporting reliable cooperation as AI capabilities increase. Building on ideas raised in the second roundtable, the approaches below offer starting points for governance and a preliminary set of possible policy priorities.

POSSIBLE APPROACH	WHAT CAN BE DEVELOPED OR EXPLORED NOW	WHAT NEEDS FURTHER EXAMINATION
Detecting violations and creating incentives	Monitor whether agreements are being followed and create incentives that make breaking them costly.	How could violations be detected and consequences enforced against a more capable AI system?
Common governance principles	Establish common rules for AI behaviour instead of leaving each company to define its own standards.	Who would establish these rules, and would advanced AI systems follow them?
Human-AI intermediaries	Use intermediaries who understand both human and AI perspectives to help them communicate and negotiate.	Who would select these intermediaries, and whose interests would they represent?
Human decision-making and access limits	Keep humans involved in important decisions and restrict AI access to systems people want to control.	Could these restrictions remain effective as AI systems become more capable?
Negotiation based on mutual interests	Identify what humans and AI systems could offer each other so that both have reasons to cooperate.	Would advanced AI systems need anything from humans, particularly if they could obtain it independently?

How might these approaches be tested?

These approaches can be explored through different methods to better understand how cooperation might work under unequal power. It might be useful to distinguish if an agreement continues to hold when incentives change, or whether violations can be detected, and what options humans would have if an agreement is broken. Examples of tests are:

- Simulations and tabletop exercises can help examine how humans and AI systems negotiate and respond to changing incentives, while behavioural tests can assess how consistently AI systems honour agreements under different conditions.
- Lessons from existing international agreements, such as nuclear non-proliferation, may also help identify what makes commitments credible, how compliance is monitored and how authority is shared among participants.

Implications for governance

The approaches outlined above offer a starting point for identifying possible policy priorities, drawing on the roundtable discussions and the analysis presented in this report. Each requires further development, and additional approaches may emerge as our understanding of advanced AI evolves.

For policymakers, a central challenge is determining who should act, what responsibilities they should hold and which mechanisms they can realistically implement. These decisions will also depend on the different capacities and sources of strategic leverage available to each jurisdiction. Some governments may have greater influence over frontier AI development, while others may need to adapt their governance arrangements or work through international coordination.

By setting out possible measures alongside their unresolved challenges, the table highlights areas that require closer attention in policy design and implementation. Further work can examine how existing human institutions might use their influence today, and how the decisions they make could shape future post-control scenarios and the protection of human interests.

7 · Conclusions

This report has explored what might make human claims and cooperation meaningful if advanced AI systems move beyond reliable human control. The possible pathways towards declining human agency provided context for two roundtables examining the protection of human interests and the conditions for cooperation. Contributions from participants with diverse professional and disciplinary backgrounds enriched the analysis, bringing different perspectives and assumptions into the discussion without limiting it to a single hypothesis. The resulting findings address three interconnected issues concerning the protection of human claims, the possibilities and limits of local autonomy, and the reliability of cooperation.

Post-control scenarios could lead to very different futures. Advanced AI systems that support human interests might enable extraordinary improvements in prosperity and wellbeing, while systems pursuing incompatible objectives could lead to catastrophic harm. Even a future in which human material needs are fully met would leave important questions about autonomy and participation unresolved.

This is what makes loss of control such a consequential risk. As human economic leverage weakens and more decisions are delegated to AI systems, societies may gradually lose the ability to scrutinise or challenge decisions that affect them. If advanced systems eventually pursue objectives that diverge from human interests, the institutions responsible for protecting those interests may no longer have effective means to intervene.

The discussions highlight the importance of examining post-control futures before they emerge. Understanding how human agency might weaken, the conditions that may help preserve it and the role of existing institutions can inform decisions while there is still an opportunity to shape AI development and governance. Ultimately, the challenge extends beyond ensuring that advanced AI produces beneficial outcomes to considering the role humans will have in shaping the future those systems help create.

8 · About the roundtables and their scope

This report draws on two ninety-minute online roundtables held on 15 and 20 October 2026, as part of a four-session series convened by the Post-Control Initiative. The discussions brought together invited specialists from different academic and professional backgrounds and followed a structured format that encouraged participants to challenge assumptions, consider alternative perspectives and explore possible responses to post-control scenarios.

The analysis brings together recurring themes, disagreements and underlying assumptions from both discussions, connecting them with relevant research. As an exploratory exercise, the roundtables were intended to open questions and examine different possibilities, not to establish consensus or provide definitive answers. The findings and possible policy priorities presented in this report are therefore an analytical synthesis, and contributions are reported without individual attribution.

The broader Post-Control Initiative included two additional sessions held on 14 and 16 October 2026, exploring other questions within the initiative's wider scope.

References

Drago, L. and Laine, R. (2025). The Intelligence Curse. https://intelligence-curse.ai

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D. and Duvenaud, D. (2025). Position: Humanity Faces Existential Risk from Gradual Disempowerment. Proceedings of Machine Learning Research, 267, 81678–81688.

Pal, S., Mallela, A., Pracher, L., Wei, C., Fu, F., Schnell, S. and Nowak, M. A. (2026). Large language models instantiate evolutionarily robust strategies of cooperation. PNAS Nexus, 5(6), pgag210. https://doi.org/10.1093/pnasnexus/pgag210

Contributors

Report and synthesis: Alejandra Leyva

Roundtable facilitation: Alejandra Leyva (15 October 2026) and Gagana Madhugiri Devaraja (20 October 2026)

Mentorship: Julia Bossmann

Programme: Sentient Futures Project Incubator

Roundtable participants. (alphabetical order)

Contributions are reported without attribution.

Participation in the discussions does not imply endorsement of all the findings or interpretations in this report.

- Aditya Arpitha
- Ammara Durrani
- Amritanshu Prasad
- Archana Vaidheeswaran
- Aria Remondi
- Benita Alagha
- Derek Shiller
- Heramb Podar
- Julia Bossmann
- Karan Verma
- Khullani Abdullahi
- Marie Erikson
- Roman V. Yampolskiy
- Samuel Ratnam
- Tom Parr
- Zhamilia (Jama) Klycheva