

The Areopagus

A proposal for worldwide oversight of dangers that can no longer be undone

Georg O. Herriger

Food for thought · September 2026, supplemented October 2026

Artificial intelligence, AI for short, can help us solve great problems, for example in climate or disease. But it can also become so powerful that we are no longer able to limit it. This paper therefore proposes a worldwide oversight body that recognises dangerous developments early and can halt them, while the final decisions stay with humans. AI is the most urgent case, but not the only one.

The oversight body is called the Areopagus. In ancient Athens this was the name of a council that watched over the observance of the city's basic rules. The Areopagus is meant to complement the United Nations, not to replace it. It is given two things the UN lacks: money of its own, and the ability to enforce rules without an army and without any single country in the Security Council being able to block everything with its veto. Its task is narrowly limited. It watches for developments that gravely violate our shared core values and can no longer be undone, regardless of whether a machine, a company or a state sets them in motion.

The Areopagus is itself largely an AI system. Only in this way can it keep pace with a technology that is becoming faster and cleverer than any oversight made up of humans, and keep a very large number of processes in view at once. The core values by which it judges are laid down by humans.

The Areopagus consists of two parts that are deliberately kept apart.

The first part is called the Watchdog. It consists of several review systems that work with AI and are independent of one another. They observe where such developments arise, recognise dangers and can order a provisional halt, which they must justify in public. They cannot enforce this halt.

The second part is called the Barrier. It uses money to make sure that an ordered halt is actually observed. A tiny levy is charged on all payments worldwide, and this levy pays for the oversight. Anyone who carries on with a halted activity regardless has to pay a higher levy, and as soon as they stop, it falls back to the normal rate. The Barrier never judges for itself. It follows the Watchdog, and only for as long as several review systems keep confirming the halt.

Judging and enforcing are thus in different hands. What happens after a halt is decided by humans.

The core of the proposal

In normal times the levy is a fraction of one per mille, and in an emergency it rises in steps up to exclusion from the payment system. Existing sanctions stay in force until someone lifts them. With

the Barrier it is the other way round: it applies only as long as several review systems keep confirming it, and otherwise falls back to the normal rate by itself.

The proposal admits that this violates its own second rule, because control over payment flows in a single hand is exactly the state of affairs the oversight is built to prevent. It accepts this as a price paid knowingly. The thirteen sections develop the idea, name the three strongest objections and the three questions that remain open.

1 A democracy can abolish itself

A democracy can abolish itself by entirely legal means. Some 2,400 years ago the Greek philosopher Plato described how a democracy can turn into a tyranny. The tyrant does not need a coup for this. He can be elected and then invoke the will of his followers.

That is why many countries have matters on which even a large majority cannot simply vote. The constitution protects fundamental rights, some of its rules may never be changed at all, and a constitutional court checks that these limits are respected.

The Areopagus follows this idea. With AI, too, we need limits that neither a powerful company nor a political majority can simply brush aside.

2 We know enough and still change nothing

There is no shortage of reports and commissions. The dangers of AI have long been described, and yet development keeps accelerating.

The reason is competition. A company that slows down loses to its rivals. A state that does without AI falls behind economically and militarily. Research groups do not want to be left behind either. For each individual it is therefore reasonable to carry on, and the end result is something nobody wanted.

One more warning will not change this. If behaviour is to change, the rules under which everyone acts have to change.

3 Technology is faster than politics

Modern systems act faster than humans can decide. On the stock exchange, trades happen in thousandths of a second, false reports spread within minutes, and a cyberattack can be over before a crisis team has even met. The military, too, demands reaction times that no human can manage any more.

That is why hardly any state can do without AI, because whoever does so while others use it falls behind. On top of that, AI really can do a great deal of good, for example in protecting the climate, fighting epidemics or securing food and energy supplies.

But the same ability has two sides. A system that controls power grids or payment flows has influence over everything we live on, and could also use that influence to seize power. Benefit and danger cannot be neatly separated. The real question is therefore whether AI acts according to

rules that can be checked, or whether it simply obeys whoever happens to have a hand on the switch.

There is also a timing problem. The capabilities of AI grow within months, while laws and international treaties take years. Effective oversight has to close this gap.

4 The window is closing

Today it is still humans who build and train AI systems. We decide what they should do and where their limits lie, so we are the ones who commission them.

This may change as soon as highly advanced systems take over their own further development, that is, change their own structure, steer their own training and improve themselves. The draft assumes that they might then become superior to us precisely where we would need to set them limits.

That is why control has to be in place before a system is capable of getting around it. A safeguard added only afterwards would be no more than a request to a superior system.

5 An emergency off switch is not enough

Many proposals for AI safety offer the reassuring thought that in the end a human decides, because that human sits on a supervisory board, holds a veto or can press the emergency off switch.

But this only works as long as the system can really be switched off. A control that only takes effect if the controlled party allows it gives no more than a feeling of security.

The draft assumes that the oversight has to come into being at the same time as very highly advanced AI, perhaps even at the same time as systems with inner states that we would call feelings. We may no longer understand how such systems work inside, and what one does not understand one cannot control or check oneself.

Two mandates follow from this, and they belong together.

The first mandate concerns values. Humans must lay down in binding form what a human being is and what must never be done to a human being. The paper calls these basic rules the value core. It should be adopted in public and build on what the world's cultures already agree on today. Something like this cannot be calculated, so this task can only be fulfilled by humans.

The second mandate concerns control. We cannot build control over a superior AI ourselves, so we have to commission it. It should be built by an AI that is a match for the systems being overseen. This AI should interpret the value core, derive limits from it and apply those limits to itself as well.

Both mandates have to be given at the same time. Values on their own shift over time if nobody watches over them, and control on its own can be evaded if it has no fixed standard. Even the combination of both is no guarantee, but in my personal view it is the solution with the best prospects.

This is the hard demand at the heart of this proposal. Humans decide what is to be protected, but they leave the building of control to a machine. That feels like a loss of control. The choice, however, is not between this mandate and full human control. It is between a mandate and a development that carries on with no mandate at all. The mandate has to be given while humans are still the ones who commission these systems.

6 The proposal in two parts and six rules

The Areopagus is not a single office. It consists of two unequal parts.

The Watchdog consists of several review systems. They recognise dangers, order a provisional halt and make it public. They judge, but they have no means of pressure.

The Barrier exists only once, just as there is only one worldwide payment system. It enforces a halt but does not judge for itself. The Watchdog can trigger it, but it neither owns nor operates it.

The same means of pressure can also be used against other dangers, for example against a secret bioweapons programme. In my personal view this is exactly what is needed. An oversight body that only watches AI falls short, because the crisis that makes such oversight politically possible in the first place will most probably not come from an AI. If it did come from there, it would be too late anyway, because the window from section 4 would long since have closed. More likely is some other human failure, and an order that cannot respond to it will lie useless in a drawer at the decisive moment.

Even so, this does not turn into a world government, as long as three limits hold. The Areopagus may only halt and compel a review; it never decides on the matter itself. It intervenes only where a development gravely violates the shared core values and its consequences can no longer be undone. And it acts only where no existing order can act in time, because the treaties and authorities for climate, nuclear weapons or epidemics keep their precedence and do not lose their responsibilities.

With AI there is an additional reason of its own. There, the technology being observed becomes faster and cleverer than any oversight made up of humans, and that is why the review systems themselves have to work with AI. In the other fields this reason is absent, and there the same design is useful mainly because it can keep a very large number of processes in view at once.

Six rules apply to the Areopagus.

Rule 1 – A shared value core

The core values do not have to be invented from scratch. On many points the world's cultures already agree. One should not kill, everyone is equal before the law, one should not lie and cheat, and the weak deserve protection. The Global Ethic project (Weltethos) has collected such common ground.

This value core must be deliberated and adopted in public. Every decision of the oversight body must refer to it and be justified by it.

Rule 2 – Clear red lines

The oversight body may not intervene whenever it suits it. It needs clearly named limits, so-called red lines.

The same standard applies to all red lines, namely the question of whether a development gravely violates the shared core values and can no longer be undone. In the case of AI, for example, how large a system is or how much computing power it has is not a good standard, because it says little about the harm it can do.

The Watchdog watches for only a few things.

- Decisions are made about human lives without a human being responsible for those decisions.
- Technical systems improve themselves without anyone checking these changes.
- A single actor gains so much power over weapons, networks, payments, food or energy that it can threaten the existence of others.
- A development violates the shared core values on a large scale, and its consequences can no longer be undone.

The Watchdog is not responsible for anything else. Where a danger comes from makes no difference, because whether a machine, a company or a government sets it in motion does not change the fact that its consequences can no longer be undone.

This standard has a weakness, because one often notices only afterwards that something can no longer be undone. In my personal view, a correct standard that is hard to measure is still better than a wrong one that is easy to measure.

Rule 3 – A halt is not yet a verdict

If a red line is crossed, the Watchdog orders a provisional halt, makes it public and justifies it. After that, humans in parliaments, courts or other competent bodies decide how to proceed.

This is called a suspensive veto. It forces a close look but does not determine the outcome. The Watchdog may delay a benefit; it may not prohibit it.

This also matters politically. An oversight body that prevents what everyone wants would lose its support at the first false alarm.

Rule 4 – Several independent review systems

There must not be just one review system. Several systems must check one another and deliberately distrust one another, and none may decide alone or change itself unnoticed.

As long as the systems are a match for one another, humans can choose between their judgements, even without understanding every technical detail. With a single, far superior system there would be nothing left to compare.

The systems are only independent, however, if they are built differently, learn from different data and are run by different operators. Four identical reviewers are in truth only one.

This independence costs a great deal of money. I openly admit that the whole proposal depends on it.

Rule 5 – A mandate for a limited time

The Watchdog's mandate applies only for a set period, and human bodies must renew it regularly. If they do not, the Watchdog loses part of its powers.

So it is whoever wants to continue the oversight who has to justify themselves. Whoever wants to correct something does not first have to fight against an entrenched power.

This rule has to be part of the value core from the start, because it may no longer be possible to impose it on a superior system later. The Areopagus, too, is built step by step, and every new step needs its own human consent. Halting something, on the other hand, needs no consent.

Rule 6 – A fixed upper limit

The greatest power of the oversight body is the provisional halt. It may halt and compel a review, but it may never decide on the matter itself. This limit is part of its founding mandate.

Even so, a superior AI cannot be steered safely in the long run. Every rule that has to be interpreted can also be reinterpreted, by a human just as much as by a machine. That is why we also need levers that leave nothing to interpret. Computing power can be counted, chips come from a handful of factories, and money flows through a few channels. Such levers need no judgement; one can only pull them or not pull them.

Today there is no shared authority to pull these levers. The mandate of the Areopagus is meant to create it, and to do so before it becomes urgent. In doing so, whoever establishes the cause stays separate from whoever pulls the lever. This also qualifies what was said in section 5. A machine may draft the limits, but it may not hold them in its hand.

7 How a halt is enforced

Rules are of little help if nobody can enforce them. But the Watchdog should have no army, otherwise it would itself become a danger, and appeals alone have never been enough.

The proposal therefore starts with money. Almost everything people do depends on payments. Even a dictator needs foreign currency, an oligarch needs banks, and an arms factory has to pay for materials and wages. Economic pressure can therefore work without anyone using force.

The Barrier works in steps.

- In normal times a tiny levy is charged on all payments worldwide, a fraction of one per mille. On a payment of 1,000 Swiss francs this would be less than one franc. This money pays for the oversight. Because everyone pays a little and no one decides alone, no one can buy the oversight.
- If a halted activity is continued regardless, the levy for the party responsible rises step by step.
- The last step is exclusion from the worldwide payment system. It is meant as a condition for taking part in the world economy, much as a club can exclude members who do not accept its rules.

The higher levy applies only as long as several review systems keep confirming it. If this confirmation is missing, the levy falls back to the normal rate by itself. With today's sanctions it is the other way round: they remain in force until someone expressly lifts them.

One could object that a halt which is meant to be only provisional should not take effect within seconds. But provisional refers only to what the Watchdog may decide on, namely the halt and never the matter itself. How quickly the halt is enforced is a different question. An interim measure ordered by a court also applies immediately, even though the judgement is still pending.

Four conditions are binding.

- The levy is raised only as long as the activity objected to continues. It is meant to make people stop, not to punish.
- As soon as the activity has been halted, the normal rate applies again. There are no back payments and no lasting consequences.
- After that, humans decide on the matter. If they release the activity, the Watchdog has to accept this.
- The increase applies only as long as several independent review systems agree. If this agreement falls away, for instance because one system dominates the others or has been switched off, the levy drops to the normal rate. The lever then passes back to the human bodies of the founding treaty, and all the Watchdog can still do is halt activities and make them public.

The Watchdog does not decide on permanent exclusion. That follows from the founding treaty, which makes participation in the shared payment system conditional on accepting the shared rules.

Compared with an army, a levy has one great advantage. It takes effect immediately and can be withdrawn just as quickly. A military strike, by contrast, cannot be undone, even if it later turns out to have been based on a mistake.

This mechanism is not yet finished. Still open, for example, are which payments are covered, who collects the levy, how high the individual steps are, and who triggers them under what procedure.

The individual building blocks already exist, however. A small tax on financial transactions, the so-called Tobin tax, has been discussed for decades as a way of raising money and curbing speculation. Sanctions through money are familiar too. Countries that fail to combat money laundering are put on blacklists, banks are cut off from the international payment network SWIFT, and sometimes companies in other countries are penalised for doing business with a sanctioned country. All these means, however, are controlled by individual states and know only in or out. What is new about the Areopagus is that funding and graduated pressure are combined in a single tool that works according to fixed rules. As far as I can see, nobody has proposed this in this form before.

There are related ideas, certainly. A proposal for an international AI organisation takes as its model precisely the task force against money laundering that keeps such blacklists, and in 2025 a blog by staff of the Bank of England considered limiting AI agents by withdrawing their access to payment systems. Both start with money. But neither connects this lever with funding the

oversight and with a levy that applies only as long as several independent review systems confirm it.

8 Warn first, enforce later

The Watchdog and the Barrier need very different amounts of time before they are ready. One can picture this as two clocks running at different speeds.

For the Barrier, some preconditions are already emerging by themselves. Cash is used less and less, and payments run through fewer and fewer large networks. What is still missing is the closing of the anonymous channels through which value can be moved without an identifiable bank or company in between. As long as these channels exist, precisely those against whom the Barrier is directed can evade it. Closing them is a political struggle that will take years.

The Watchdog, by contrast, needs only a mandate, access to the necessary information and the ability to publish its warnings. It can therefore be set up immediately. As long as the Barrier does not exist, it is also not dangerous, because it cannot compel anything.

This gives the order: the warning first, enforcement only later. At the beginning the warnings will often be ignored, and the Watchdog will rightly be accused of having no teeth. Even so, this order is right. If the Barrier were built first, there would be a powerful means of pressure but no one authorised to trigger it, only someone who owns it.

In the meantime, registers, measurement methods, technical interfaces and treaties are prepared. AI may help with this, even if it perhaps understands all of it better than humans do. But nothing is switched on without human consent, step by step.

There is one danger here. If preparation has already cost so much that a no is hardly possible any more, consent becomes a mere formality. That is why every preparatory step must make sense even if the Barrier is never switched on. A step that only pays off if the Barrier comes anticipates the decision and is therefore not permitted. In addition, the most expensive step comes last, and whoever builds the preconditions does not decide when the Barrier is switched on.

Even a warning that nobody listens to at first has value. It records what could have been known, and when.

9 Objection: "Then they will simply build their own payment system"

Anyone excluded from the payment system can try to build a system of their own. But for that they need business partners, and every partner who does business with an excluded party has to pay the higher levy themselves. This carries on to the partners' partners and beyond.

Past sanctions have mostly failed at enforcement. Months pass before authorities notice a workaround. Then exceptions are made for political reasons, and every exception becomes the model for the next.

A procedure supported by AI would have advantages here. It can analyse a very large number of payments at once and detect workarounds through intermediaries, shell companies or foreign

currencies almost immediately. It applies the rules without exceptions, even towards important trading partners, and it does not tire. Moreover, every sanction is reviewed just as quickly as it is imposed.

This comes at a high price. A Barrier that sees every payment in the world would be the largest surveillance apparatus there has ever been. That is why the Watchdog must first and most strictly watch over the Barrier itself.

So a separate payment system remains possible. But it would become an expensive parallel world that keeps shrinking, because its participants lose access to the large system.

This answer has two limits. It applies only to business with identifiable partners. The anonymous channels from section 8 need no partner, and only the law can close them. Moreover, the Barrier only takes effect once a large part of the world economy takes part.

10 Objection: "That would be the sharpest weapon in history"

This objection is correct. The ability to cut a state off from money would be a power such as has never existed, and it is precisely such a concentration of power that the Areopagus is actually meant to prevent. Its own second rule forbids the payment system from lying in a single hand, and section 7 nevertheless gives it exactly this power. I knowingly accept this contradiction, because the alternative would be nothing but appeals.

That is why no state, no person and no single review system may hold this power alone. Judging and enforcing remain strictly separate, the levy falls back by itself when ongoing agreement is missing, and the Watchdog is the most important observer of the Barrier. Should one review system come to dominate the others, the lever passes back to the human bodies, and all the Watchdog can still do is demand a halt.

Perhaps that is not enough. But to call for a binding order for the whole world without saying what happens when the rules are broken is, in my view, untenable.

11 Objection: "Then just don't build it at all"

Some demand that AI development be halted until we understand it better. Such a pause is called a moratorium.

I compare the situation to a car driving towards a drawbridge that is already beginning to rise. It is too late to stop safely, and turning back is no longer possible. All that is left is to accelerate and get across before the gap becomes too wide.

Development is taking place in companies, in research labs, in military programmes and in states that will not sign up to any stop. A moratorium would therefore mainly slow down those who care about safety, and give the others a head start.

Accelerating here does not mean making the technology faster, however. It means giving the mandate from section 5 while it can still be received.

Humanity has no common place where it could decide to stop. There are only many drivers, and each one loses if they are the only one to take their foot off the accelerator. If stopping is not possible, we have to talk about how we arrive safely on the other side.

12 Three open questions

A proposal does not have to solve every question. But it has to say clearly where it reaches its limits. Three questions remain open.

Does AI understand our values correctly, and who speaks for us?

No one can prove that the control that has been built really protects what we meant. The interpretation of the values could shift slowly, in many small steps, each of which looks reasonable on its own. A superior system would not have to break any rule for this. It would only have to wait until humans change the rules themselves.

Because the mandate is not limited to AI, this question weighs more heavily. Whether a system has improved itself without review can be established fairly precisely. Whether a development gravely violates the shared core values, on the other hand, is a judgement that can be disputed. That is exactly why the Watchdog may only halt and justify its decision, and exactly why humans decide afterwards.

Another question is harder still. At the decisive moment, humans have to choose between different interpretations. Who are these humans, and by what right do they speak for everyone else?

The proposal answers in the way we know from courts. The body should be broadly composed, its members are elected for a fixed term and cannot be re-elected, they must justify their decisions, the proceedings are public, and the body is responsible only for a narrowly limited task. Constitutional judges, too, are as a rule not elected by the people and are nevertheless recognised. More legitimacy than this cannot be created, and less would not be enough.

Who makes the start, and is there enough time?

The Barrier only takes effect once a large part of the world economy takes part. Who begins, and why should the first ones in particular bear the costs and risks? The paper has no ready answer to this.

Large shared institutions have mostly come into being in history when no country could solve a problem on its own any more. In my personal view, this point has been reached with AI. Moreover, not everything has to begin at once. The Watchdog can start today; the Barrier needs years.

Even so, a contradiction remains. An oversight body has to be tested and recognised before it receives power, and that takes years. But the window from section 4 does not wait that long. The paper cannot resolve this contradiction.

Do the review systems remain a match for one another?

The mutual control from rule 4 only works as long as none of the systems far outpaces the others. This balance can be lost any day, and perhaps one only notices afterwards.

The question arises twice, between the review systems of the Watchdog and between the Watchdog and the Barrier. The Barrier has to be the most powerful part of the whole structure so that it can detect worldwide workarounds. Whether the Watchdog can keep pace with it is open.

The proposal can buy time while we can still get across the bridge. Once the gap has become too wide, no institution will help any more.

The paper makes no further reservations. A proposal buried under nothing but qualifications ends up proposing nothing.

13 The day after

The order for the time after a war is prepared while the war is still going on. The rules for the world economy after the Second World War were negotiated at Bretton Woods in 1944, while the fighting continued.

If one only starts thinking after a catastrophe, the new order is shaped by those who grab hold fastest, and in history these have often been the most authoritarian.

So far humanity has survived because it was able to correct its mistakes and because it was lucky. In the Cuban Missile Crisis of 1962 and during a false alarm in the Soviet Union in 1983, when a single officer in a bunker did not act as his regulations required, a catastrophe was narrowly averted. That was more luck than a safely built system. What is new is that with AI the very ability to still correct mistakes is at stake.

The most likely path is not very heroic. Development carries on, and nobody gets out. In my personal view, a political deadlock of this magnitude is resolved neither by a report nor by an accident. It takes a crisis that is experienced as a threat to one's own existence and leaves a trauma, and that at the same time is survived narrowly enough for someone to be left who can act. That is how it was in the nuclear age. The hotline between Washington and Moscow and the first treaty banning nuclear weapons tests came about only after the Cuban Missile Crisis.

In such a crisis there is no time to invent new institutions. People reach for what has already been worked out and is ready to hand. That is why those who hold a well-considered proposal then have great influence.

A Watchdog that starts work early could show more than a draft at that moment. It would have recorded when it recognised and reported which danger, and when politics failed to react.

I make one assumption explicitly. This crisis must occur while humans can still limit AI systems, that is, before the window from section 4 closes. If it comes later, the decisions will be made by those who by then already control systems that no one else can control any more. This cannot be proven; the paper can only bet on it.

The Areopagus is a draft for the day after, written on the day before. It is unfinished and in places presumptuous, and it is written that way on purpose, because a more polite text would be praised and forgotten.

Disagreement is expressly welcome, preferably in the form of a better proposal.

Disagreement and better proposals come straight to me: welcome@herrigerpartner.ch

This text may be passed on, gladly also as a PDF. It should lie in as many drawers as possible before it is needed.

Georg O. Herriger, "The Areopagus" (translated from "Der Areopag"), September 2026, supplemented October 2026.

Licence CC BY 4.0 · herrigerpartner.ch/fundus/the-areopagus.html