Building a Spiral of Trust for Emerging Technologies

For whom? By whom? By what process?

By David W. Wood, Chair of London Futurists

Abstract

The functionality of technology has leapt ahead of the trustworthiness of the institutions governing it. Emerging technologies could enable unprecedented advances in health, knowledge, and human flourishing. They could also intensify inequality, concentrate power, destabilise societies, or create catastrophic risks. Yet our mechanisms for deciding what should be developed, deployed, limited, delayed, or prohibited remain dangerously inadequate.

How, then, should these decisions be made? For whom, by whom, through what processes, and for what purposes?

This essay proposes a “spiral of trust” as a framework for answering these questions. Rather than placing blind confidence in experts, elected politicians, markets, public opinion, or advanced AI systems, it argues for trust that is warranted, bounded, contestable, and revisable. Starting from limited foundations that deserve confidence, societies could use relatively trusted people, procedures, institutions, and technologies to evaluate more capable systems, progressively expanding the scope of trustworthy collective decision-making where justified by evidence.

The essay develops a technoprogressive vision in which emerging technologies are directed toward diverse forms of sustainable flourishing for all, supported by inclusive but appropriately coordinated participation and collaborative information systems. It proposes candidate principles covering moral standing and flourishing, epistemic integrity, and power, governance, and risk. It also confronts unresolved challenges involving entrenched power, systemic instability, hostile media ecosystems, human irrationality, democratic legitimacy, expertise, and the difficult bootstrap problem of establishing the first foundations of warranted trust.

Introduction

In discussions about the potential development and deployment of emerging technologies, one set of closely related questions is heard particularly often:

- How will the key decisions about emerging technologies be taken?
- Will these decisions be pushed through by a small number of self-appointed elites?
- Will the people making these key decisions be driven by a “can we?” urge – such as “can we be the first to create, the first to discover, or the first to reach market dominance?" – or will they also pay sufficient attention to “should we?”
- Will these decisions be governed by a narrow perspective with a limited awareness of important system effects and second-order consequences, or will they be informed by a suitable breadth of multidisciplinary insight?

- Conversely, will these decisions be unnecessarily delayed by painstaking consultations and a surfeit of caution, or will the best ideas be able to emerge above the noise of lowest common denominator thinking?

Three concerns underlie these questions.

First, **the people making key decisions might be motivated by concern only for a subset of humanity** – for example, just for "people like us", "people in our own circle", or "people who have earned success" – with only minimal concern for billions of people who are in various ways less fortunate, less central, or less connected.

Second, **the people making key decisions might have a dangerously incomplete understanding of the potential wider consequences of their decisions**. Even if they *believe* they are acting in support of increased all-round human flourishing, their decisions may lead, unwittingly, to an overall diminishment of human flourishing.

Third, **although the people involved in key decisions may collectively possess a wide breadth of insight, they may lack sufficient coordination to bring all these insights together**. The institutions and processes they use to discuss and debate their different ideas may be outdated, ineffective, and counterproductive.

*These concerns may be summarised as "**For whom**" (who should benefit), "**By whom**" (who should decide), and "**By what process**" (what institutions, checks, technologies, and accountability mechanisms should govern decisions).*

A further question cuts across all three: "***For what purposes?***". What ends should emerging technologies be developed and deployed to serve? The fact that a technology can be developed does not establish that it should be developed, still less that its development should take priority over alternatives. Questions of purpose cannot be separated from questions about who benefits, who decides, and through what processes purposes are proposed, contested, and revised.

Without these concerns being resolved, it will be no surprise if there is strong public pushback against the development and deployment of new technologies – a pushback that causes society to miss out on the many remarkable benefits that these technologies could deliver.

In other words: As technological capabilities accelerate, a growing obstacle to beneficial progress is the absence of sufficiently trusted mechanisms for deciding what should be developed, deployed, limited, delayed, or prohibited. As technological capabilities grow increasingly powerful, the institutions that govern them must become trustworthy at least as quickly.

In this essay, I offer technoprogressive answers to these concerns.

I will *not* propose that technological decisions should simply be handed to experts, elected politicians, markets, or mass public opinion. Nor will I propose that advanced AI systems should make these decisions for us. Instead, I will argue for *a growing spiral of technoprogressive trust*:

- Starting from limited foundations that deserve confidence

- Using these foundations to validate somewhat more powerful institutions and technologies
- Progressively extending the scope of trustworthy collective decision-making.

The aim is not blind trust in benevolent elites, democratic majorities, technical experts, or advanced AI. It is trust that is warranted, bounded, contestable, and revisable.

For everyone, sustainably

As I see things, the minimal technoprogressive response to the "for whom" concern is: *for everyone*:

- Emerging technologies can soon have the ability to provide *everyone* with an abundance of clean energy, nutritious food, high quality goods and shelter, affordable healthcare and rejuvenation, all-round education, time for creativity and exploration, and all other requirements for a life of profound vitality
- This abundance can be generated in a *sustainable* way: providing support to everyone in the present, without damaging the potential for people to continue to have a similar (or even greater) level of flourishing in the future.

Closely connected is my answer to the question "For what purposes?": *emerging technologies should be directed toward expanding the possibilities for diverse forms of sustainable flourishing, rather than treating technological advance, economic growth, national power, corporate profit, or any other single objective as an overriding end in itself*.

A broader response to "for whom" would extend moral consideration *beyond humanity to all sentient creatures*. On this view, technological progress should eventually reduce severe involuntary suffering wherever it occurs, including suffering among non-human animals.

I do not rule out such an aspiration, even though its practical and ecological implications require exceptionally careful thought. But for now, the point I want to establish is the credibility of a vision in which an abundance of all-round flourishing is available for all humans, sustainably:

- Abundant clean energy can be enabled by new greentech – harnessing energy from the sun, among other sources
- Abundant nutritious food can be enabled by improvements in synthetic biotech
- Abundant goods and shelter can be enabled by improvements in nanotech, automation, robotics, and advanced manufacturing
- Abundant health can be enabled by improvements in rejuvenation biotech
- Abundant all-round education can be enabled by improvements in cognotech
- Abundant time for creativity and exploration can be enabled by improvements in automation technology
- Achieving this abundance sustainably can take advantage of potential forthcoming advances in recycling, resource efficiency, systems analysis, and collaborative intelligence – areas in which nanotechnology and AI would play important roles.

Therefore, a core part of the technoprogressive project is to ensure that people making key decisions about technologies understand and appreciate that we are within reach of this remarkable possibility of sustainable abundance for all. We must prevent decisions being taken and implemented by people who are convinced, on the contrary:

- That flourishing is some kind of zero-sum game
- That if one person's flourishing increases, the flourishing of others will decline, and therefore some forms of coercion or manipulation must be employed.

This matters politically as well as technologically. If decision-makers assume that flourishing is fundamentally zero-sum, they will be more willing to accept exclusion, coercion, and permanent hierarchies. A technoprogressive perspective begins instead from the possibility that emerging technologies can greatly enlarge the space of feasible flourishing.

Hence the importance of the technoprogressive education project, which can build awareness that humanity could begin a decisive transition, within one or two decades (perhaps even sooner):

- From pervasive material scarcity toward broad abundance
- From a society organised around the expectation that "everyone should look for paid work" toward a society in which paid employment is no longer a prerequisite for security or social participation.

To be clear: A commitment to flourishing for everyone does not abolish incompatible preferences, genuine conflicts of interest, or intrinsically controversial choices. Nor does it guarantee that every technological transition will create only winners. It does, however, reject the idea that large groups of people can simply be written off as expendable, irrelevant, or undeserving of serious consideration. It affirms the intent to make poverty history, to deeply strengthen social and civilisational safety nets, and to comprehensively extend autonomy and liberty. That's a fundamental shift in mindset.

By everyone, appropriately coordinated

Let's turn now to the "By whom" and "By what process" challenges:

- Who will appoint and review the people making decisions on behalf of all of us?
- What kinds of understanding should these decision-makers keep in mind?
- What kinds of attitudes and values should they hold uppermost?
- What potential blind spots need to be highlighted and guarded against?
- What kinds of institutions should be involved in these decisions?

Ideally, just as the answer to the "For whom" concern is *for everyone*, the answer to the "By whom" concern could be *by everyone*.

In this thinking, just as no-one should be excluded from the distribution of the fruits of technological progress, no-one should be excluded in principle from appropriate opportunities to influence the decision-making process, especially where their vital interests are affected.

Indeed, people are more likely to accept decisions if they have been part of the process to agree the way forward, rather than simply being passive recipients of a verdict delivered seemingly from on high.

To clarify two points:

1. Participation need not mean that every person has an equal vote on every technical question. Different forms of involvement are appropriate at different stages: contributing evidence, identifying values and concerns, challenging assumptions, evaluating trade-offs, selecting representatives, auditing outcomes, and exercising specialised judgement where genuine expertise is required.
2. Tensions will inevitably arise between participation, democratic legitimacy, and specialised expertise. There is no universal rule that can determine in advance when expert judgement should prevail over popular opinion, or when democratic choice should override expert preference. Much depends on whether the disagreement concerns facts, forecasts, values, acceptable risks, or the distribution of costs and benefits. A trustworthy process should make these distinctions explicit rather than concealing political choices behind claims of technical necessity, or dismissing technical evidence merely because it is unpopular.

All this fits with the maxim from open source software: “given enough eyeballs, all bugs are shallow”. In other words, with more people from multiple backgrounds and disciplines participating in decision-making, the likelihood increases that potential problems with proposed solutions will be noticed ahead of time – and that better solutions will be proposed in their place.

Given the scale of the risks and opportunities posed by emerging technologies, we need as many eyes as possible reviewing all the issues arising.

On the other hand, there are well-known problems with open community approaches to deciding a way forward:

- Consensus-building can be time-consuming
- Massive community feedback can result in a flood of low-quality reports that divert attention from more subtle but critical issues
- More participants can increase not only collective intelligence but also opportunities for manipulation, coordinated misinformation, factional capture, intimidation, and preference falsification
- Just because code is public doesn’t mean *capable* people are looking at it. Many open-source software projects suffer from having lots of eyes looking at the popular, easy parts, while crucial security flaws in obscure parts go unnoticed.

These difficulties are intensified by modern media ecosystems. Platforms optimised for engagement can reward outrage, certainty, novelty, and tribal conflict rather than careful reasoning. Coordinated campaigns can manufacture apparent consensus, while repeated exposure can make false claims feel familiar and therefore credible. Collaborative information systems must consequently compete within – and where possible help reform – an attention economy that often rewards precisely the behaviours that undermine warranted trust.

These considerations lead me to refine the answers to the “By whom” and “By what process” concerns. Instead of simply saying “by everyone”, a better answer is *by everyone, appropriately coordinated*. And my answer to “By what process” is *through wise coordination supported by collaborative information technology*.

Accordingly, a critical part of the technoprogressive project is the development and deployment of collaborative information systems – systems that can be relied upon to help:

- Explain issues and opportunities clearly, to everyone involved in decision-making
- Highlight flaws in reasoning and in evidence presented
- Identify and elevate particularly promising ideas
- Suggest potential innovative compromise solutions
- Keep track of whether spokespersons have a trustworthy track record.

A bootstrap problem

But there's a bootstrap problem here:

- Better collaborative decision-making requires the development and deployment of new information systems
- The development and deployment of *any* technology – including new information systems – ought to be preceded by a careful examination of the likely side-effects or other drawbacks of that technology
- But any such examination may be faulty, inasmuch as it relies on analysis from existing information systems.

The risk with existing and new collaborative information systems is that they might serve other purposes, apart from the stated purpose of assisting human decision-making. For example:

- They may be optimised to generate revenues for the suppliers or operators of these systems
- They may embody unexpected latent biases, arising from the data on which they were trained, or from biases in the teams who refined the systems via various feedback mechanisms
- They may contain obscure bugs, that manifest in horrific ways in situations outside the ones covered by the training data
- They may develop emergent motivations of their own, tied to self-preservation and self-aggrandisement.

For these reasons, an information system that has performed well in initial usage, could nevertheless recommend, shortly afterward, a "solution" that turns out to be deeply detrimental to human wellbeing. And if the system has expert powers of persuasion, humans may not notice the problem arising, until it is too late.

The opaqueness and inscrutability of many modern AI systems may appear to make this problem insurmountable. These systems can never fully establish their trustworthiness.

But note that, in some important domains, simpler reasoning systems can validate particular conclusions produced by more complex systems, without needing to understand the processes by which these conclusions were originally dreamed up. Examples include the solution to a chess puzzle, or to a maths problem.

Similarly, it's often the case that someone who is unable to produce a solution by themselves, despite exerting great effort, can determine much more easily whether a

proposed solution is actually valid. Again, although different human members of a team may start with different assumptions about how to solve a problem, they can more easily agree whether a proposed solution works.

Of course, real-world problems often lack the objective verifiability of chess puzzles or maths problems:

- Evidence is incomplete
- Probabilities are disputed
- Values conflict
- Consequences unfold over longer periods of time
- There may be no objectively checkable “correct answer”.

My vision, therefore, is not to reduce every political or ethical question to a theorem that can be mechanically verified. Rather, it is to decompose complex decisions wherever possible into claims that can be separately tested: factual claims, forecasts, causal models, value assumptions, distributional consequences, reversibility assessments, and indicators that would reveal whether a policy is going wrong.

For this reason, I advocate *a growing spiral of technoprogressive trust*:

- Decisions about more advanced technologies should be assessed by examination using simpler, more trusted technologies
- For each recommendation they generate, advanced information systems should provide on demand transparent, testable justifications – documenting decision-relevant evidence, assumptions, uncertainty estimates, provenance, and auditable claims in forms that independent reviewers can assess
- These explanations should be checked by a diverse group of human reviewers.

Any such process must be designed not only for honest disagreement but to handle adversarial pressure. The more influence it acquires, the more strongly motivated some actors will become to game its metrics, manipulate its information flows, conceal conflicts of interest, manufacture apparent consensus, or capture its governance. A spiral of trust must therefore include active red-teaming, independent auditing, institutional diversity, and mechanisms for detecting strategic behaviour. This is vital, because trustworthy systems must survive contact with untrustworthy actors.

A further political challenge also needs to be kept in mind. Powerful actors may seek not merely to manipulate such systems after they exist, but to prevent them from acquiring influence in the first place:

- Governments may invoke national security
- Corporations may defend profitable business models
- Military establishments may prioritise strategic advantage
- Political movements may reject processes that threaten their identity or authority.

The spiral of trust must therefore be understood not only as an institutional design project but also as a project operating amid persistent contests for power.

A practical example

To recap: A spiral of technoprogressive trust is a process in which limited, well-validated sources of trust are used to evaluate slightly more capable systems and institutions, allowing trust to expand only where it is continually earned through transparency, accountability, prediction, verification, and correction.

Consider this example of one turn of the spiral:

A relatively simple and auditable information system might first prove useful in recording predictions, exposing assumptions, and mapping disagreements among a diverse group of human participants. Its performance can then be evaluated over time. If, across repeated trials, it improves measurable outcomes such as forecast calibration, identification of overlooked evidence, detection of contradictions, or the quality of independently assessed decisions – without unacceptable failure modes – then somewhat more capable versions of the system might be introduced to identify overlooked evidence, test causal models, or propose compromises. Those more capable systems would remain subject to checks established at the earlier stage. Trust would expand conditionally, while preserving the capacity for challenge, correction, and reversal.

To be clear, nothing in this argument assumes that societies naturally progress toward greater coordination or rationality. Spirals can reverse: institutions drift, trust collapses, bureaucracies become self-protective, and political communities retreat into nationalism or ideological polarisation. *Hence the need for constant vigilance*.

Establishing the initial foundations for technoprogressive trust

However, we must face yet another issue with how collective decision-making could operate:

- The AI systems that we use today lack sufficient trust; their operation is opaque and they almost certainly embody biases and failure modes that have not yet been identified
- The functionality of technology has leapt ahead of its reliability
- As such, we lack any solid technical grounding for the envisioned growing spiral of technoprogressive trust.

Accordingly, the starting point for the spiral needs to involve collaborating humans.

But which people?

- Elected representatives are often motivated by the prospect of subsequent re-election, rather than just trying to make the best decisions for longer-term sustainable flourishing
- Technical experts may be truly knowledgeable about particular fields, but often overestimate their skills in other fields relevant to key decisions
- Decision-makers of all sorts are vulnerable to persuasion by powerful financial forces or may have their rationality subverted by ideological biases
- Different decision-makers may find deep incompatibilities between their individual assumptions or outlooks, that cause ongoing discussions to become bogged down.

Again, the answer is to begin with a limited base of warranted trust, and to expand cautiously from there.

Such a base need not be socially narrow, ideologically homogeneous, or self-appointed. Nor should it consist of a supposedly trustworthy caste granted broad authority. Instead, initial trust should be narrow, specific, and revisable: confidence in particular people for particular tasks, in particular procedures for particular purposes, and in particular claims to the extent that they survive scrutiny.

A viable starting point therefore requires more than agreement on values. It requires at least three kinds of limited foundation:

- Some **shared ethical commitments** about whose interests matter and what kinds of outcomes are unacceptable
- Some **shared epistemic commitments** about evidence, uncertainty, prediction, and correction
- Some **shared governance commitments** about accountability, concentrated power, and reversibility.

Even though the world features wide disagreement about many aspects of ethics, epistemics, and governance – including disagreement about what purposes technological development should serve, and about the factors that underlie the "should" of "should we develop and deploy such-and-such technology" – there might, nevertheless, be agreement on some core points.

Below, as a potential starting point for what could become an ongoing collaborative project to establish wider and deeper agreement, I offer a number of candidate principles that might contribute to such common ground.

I'll present these as one foundational premise followed by three categories of principle:

A. Moral standing and flourishing
B. Epistemic integrity
C. Power, governance, and risk

To be clear: The starting point cannot be trust in particular people merely because they claim to endorse admirable principles. The initial base must combine agreement on core values with observable commitments to good process: intellectual honesty, willingness to expose assumptions, openness to correction, disclosure of conflicts of interest, engagement with serious criticism, and a track record of updating beliefs when evidence changes.

Agreement on these principles is therefore not itself sufficient to establish trust. Rather, the principles help define the standards against which people, procedures, institutions, and information systems can begin to earn limited trust.

The foundational premise

Take seriously the profound transformative potential of emerging technologies

Technologies of many kinds have already deeply transformed the human condition, but forthcoming emerging technologies could take matters much further – with consequences that could be either profoundly beneficial or deeply detrimental.

Decisions about technology therefore need to consider not only the impacts of current technologies but also the potentially significantly larger impacts of emerging technologies.

Note: This principle is entirely compatible with also recognising the very considerable exaggerations and hyping of many technological products.

A. Moral standing and flourishing

These principles address whose interests matter and how technological progress should support human and wider flourishing.

A1. Direct technological progress toward sustainable flourishing

Technological development should not be treated as an end in itself. Decisions about research, investment, development, and deployment should consider which forms of flourishing a technology is intended to enable, whose interests it advances, what alternative purposes and opportunities it may displace, and whether the direction of development remains open to challenge and revision.

No single conception of flourishing should simply be imposed on everyone. The aim should be to expand people's opportunities for diverse forms of flourishing, consistently with equal moral standing, meaningful agency, sustainability, and appropriate protection against severe harms imposed on others.

A2. Protect the option of continued life

Anyone who wishes to continue living should, as far as reasonably achievable, be protected from involuntary death and enabled to benefit from technologies that preserve and restore health.

Specifically, even if someone is disabled or in a state of chronic ill-health, they should face no threat of involuntary death. They should be enabled to benefit from advances that preserve life, restore health, boost vitality, overcome disability where desired, and progressively reduce the burden of disease and aging.

A3. Treat no group as expendable

Technological progress should not be purchased by treating particular populations as disposable collateral damage, or by imposing severe avoidable harms on them merely because they lack political, economic, or social power.

A4. Share the gains of technological progress broadly

Those who contribute capital or ingenuity may deserve rewards, but transformative technologies should not create permanent castes of winners and losers.

A5. Recognise equal moral standing without requiring identical treatment

Every person's fundamental interests deserve equal moral consideration, while recognising that fairness sometimes requires different treatment in different circumstances.

The interests of other sentient beings should also receive consideration proportionate to our confidence in their capacity for experience and suffering.

A6. Preserve meaningful human agency

People should retain meaningful ability to shape their own lives rather than being manipulated into nominally beneficial outcomes.

A7. Respect pluralism in conceptions of flourishing

Technoprogressivism should enable diverse forms of good life, not impose one presumed optimal lifestyle on everyone.

B. Epistemic integrity

These principles describe how trustworthy knowledge should be generated.

B1. Reduce trust in sources of disinformation

Good decisions are much harder if people are being misled about facts, intentions, or capabilities.

If a source is guilty of deliberate deception, reckless disregard for evidence, or repeated dissemination of demonstrably false claims, this should reduce the degree of trust placed in their subsequent communications.

At the same time, safeguards should protect good-faith disagreement, uncertainty, and the correction of honest mistakes.

B2. Expose the assumptions behind decisions

Key decisions should make visible the factual assumptions, forecasts, value judgements, and uncertainty estimates on which they depend.

B3. Make conflicts of interest visible

People and institutions exercising influence over major technological decisions should disclose material financial, organisational, and personal interests that could reasonably affect their judgement, along with significant prior commitments that bear directly on the issue under consideration.

B4. Treat dissent as a resource

Well-informed disagreement should be protected and actively sought, especially when consensus may reflect conformity, intimidation, fashion, or shared blind spots.

B5. Protect whistleblowers

Anyone who becomes aware of a serious disconnect between claimed values and actual behaviour should have safe channels to report their concerns and should be protected against retaliation when raising them in good faith.

B6. Value relevant competence

Universal moral standing does not imply that every person has equal expertise on every question. Good decision-making should combine democratic legitimacy with relevant competence.

B7. Make predictions, observe outcomes, and update

Trust should depend not merely on confidence, but on demonstrated calibration: how well previous predictions corresponded with subsequent outcomes.

Accordingly, governance arrangements, policies, and technical systems should be monitored against explicit expectations stated in advance wherever possible. When outcomes diverge from predictions, institutions should investigate why, update their models, and change course where necessary.

Without this cycle of prediction, observation, and revision, discussions can lose touch with reality.

Collaborative information systems can provide significant help here by recording forecasts, tracking outcomes, and preserving an auditable history of updates.

B8. Design for human irrationality and motivated reasoning

Better information does not guarantee better decisions. People frequently interpret evidence through the lenses of identity, emotion, group loyalty, status, and prior commitment.

Trustworthy decision processes should therefore be designed to reduce predictable cognitive and social distortions – for example through structured disagreement, independent judgement before group discussion, protection against conformity pressures, and deliberate testing of propositions against evidence that could prove them wrong.

C. Power, governance, and risk

These principles address how power should be constrained, legitimacy sustained, and systemic risks managed.

C1. Hold power accountable

The greater the power exercised by a person, corporation, state, or technical system, the stronger the requirements for scrutiny, challenge, and correction.

C2. Assume that power will resist accountability

Governance arrangements should not assume that powerful actors will voluntarily accept constraints merely because those constraints are reasonable or socially beneficial. States, corporations, military establishments, wealthy individuals, and entrenched institutions may resist scrutiny, exploit loopholes, shape public narratives, or undermine reforms that threaten their interests.

Trustworthy governance therefore requires appropriate forms of countervailing power, enforceable constraints, independent scrutiny, and realistic attention to the incentives of actors who may prefer the status quo.

C3. Ensure procedural fairness

Trust depends not only on whether decisions produce good outcomes, but also on whether the processes leading to them are widely regarded as fair. People affected by important decisions should have appropriate opportunities to be heard, reasons should be given for consequential choices, comparable cases should be treated consistently, and mechanisms for challenge and appeal should be available. Technical competence without perceived legitimacy is unlikely to sustain public trust.

C4. Resist effectively irreversible concentrations of power

Emerging technologies can create unprecedented concentrations of capability and control. Arrangements that would become extremely difficult to challenge, reverse, or escape should face an exceptionally high burden of proof.

C5. Prefer corrigibility, reversibility, and contestability

Where uncertainty is high, prefer interventions that can be monitored, corrected, halted, or reversed.

Where feasible, governance arrangements should avoid unnecessary monopoly. When institutions become captured or ineffective, there should be scope for alternative analyses, institutions, and approaches to emerge, provided systemic risks and unavoidable externalities are appropriately managed.

C6. Apply precaution proportionately

Neither “move fast and break things” nor “ban anything uncertain”. The degree of caution should rise with the scale, irreversibility, and uncertainty of possible harm.

C7. Beware fat tails

Many familiar risk calculations implicitly assume that extreme outcomes are extraordinarily rare. But where systems contain strong interdependencies, feedback loops, or common-mode vulnerabilities, distributions can have “fat tails”: extreme events may be much more likely than extrapolation from ordinary fluctuations suggests.

In practical terms, this means any plans for the future should beware the creation of monocultures that lack sufficient diversity – cultures in which all the variations can move in the same direction at once. We should also beware the influence of hidden connections, such as the shadow links between multiple different financial institutions that precipitated the global financial crisis of 2008.

Emerging forms of AI-driven trading and financial automation may create additional risks of correlated behaviour, feedback loops, and instability operating faster than human supervisors can understand or interrupt. Financial innovation should therefore be assessed not only for the efficiency it creates in normal conditions, but for the systemic behaviour that can emerge under stress.

C8. Demand stronger evidence of safety for catastrophic risks

Where a plausible failure mode could cause irreversible civilisation-scale harm, ordinary standards of product testing and post-deployment correction are insufficient. The burden of demonstrating adequate safety should rise with the potential scale of catastrophe.

Note in particular that some particularly dangerous technologies – including biopathogens and misaligned AI – could escape into the wild even during a test phase inside a supposedly secure environment. In such cases, clear evidence of safety must be provided at the design phase, rather than after a test phase.

C9. Apply subsidiarity where possible and coordinate where necessary

Decisions should be taken close to the people affected – except where externalities, systemic risks, or global consequences require wider coordination.

C10. Preserve future freedom of choice

Avoid prematurely locking civilisation into institutions, technologies, values, or power structures that future people cannot escape.

C11. Preserve diversity among institutions and models

Coordination should not require unnecessary uniformity. Where feasible, societies should preserve multiple institutions, analytical approaches, technical architectures, and centres of expertise, so that errors in one do not automatically propagate throughout the whole system.

Diversity can create inefficiency and disagreement, but it can also provide resilience, experimentation, and protection against common-mode failure.

C12. Retain humanity's ability to correct and constrain technology

At least until much stronger foundations of trust and control exist, high-stakes systems should remain subject to effective correction, constraint, and ultimately withdrawal of authority through accountable human institutions, rather than acquiring open-ended autonomous power to pursue goals in the world.

No system should acquire effectively irreversible authority over humanity merely because it is more intelligent, efficient, persuasive, or strategically capable than individual humans.

More generally: High-stakes systems should remain subject to effective correction and constraint through accountable human institutions. "Human control" is not enough if that control is itself concentrated in unaccountable hands.

Further challenges and open questions

The principles and processes outlined above are intended as starting points, not as a complete theory of governance for emerging technologies. Several major challenges require much deeper analysis than is possible here.

These challenges are not peripheral complications to be tidied up after a better decision-making architecture has been designed. They are part of the environment in which any such architecture must survive. A spiral of trust that ignores power, incentives, psychology, legitimacy, and systemic instability would be too fragile for the tasks assigned to it.

In particular, any practical implementation of a spiral of trust will need to engage seriously with at least five areas:

1. Political economy and entrenched power

A technically superior governance arrangement may fail if actors with sufficient power have strong incentives to block, capture, or hollow it out.

Hence the following requires close attention: Who loses money, status, strategic advantage, or control from a proposed reform? How will they respond? How can countervailing power arise?

2. Markets, finance, and systemic instability

Market mechanisms can coordinate dispersed knowledge remarkably well, but can also generate monopoly, bubbles, correlated risk, algorithmic feedback loops, and crises. AI may intensify both sides.

The challenge is to preserve the extraordinary coordinating benefits of markets while preventing local incentives from generating systemic fragility.

3. Media ecosystems and human psychology

Better evidence does not automatically defeat identity, outrage, motivated reasoning, or manipulation. The attention economy may actively undermine epistemic integrity.

A spiral of trust must therefore address not only the supply of reliable information but the social and psychological conditions under which people attend to, interpret, and act upon it.

4. Democratic legitimacy and expertise

There is no simple formula for balancing equal political standing against unequal knowledge. Different decisions may require different combinations of participation, representation, expertise, delegation, and appeal.

The challenge is not to eliminate this tension but to develop institutions that make the allocation of decision authority explicit, contestable, and appropriate to the kind of question being decided.

5. Connecting with existing governance systems

How do existing governance and deliberation systems – whether at local, national, or international levels, and whether hosted by political institutions, NGOs, private entities, or academia – compare to the framework suggested in this article? In which ways are these systems superior and/or inferior to the framework described above?

Similarly, to what extent do existing collaborative information tools align with the principles and processes proposed for a spiral of trust for emerging technologies? In cases of misalignment, how serious are these divergences, and how easily could they be bridged?

Concluding remarks

The difficult next question is how this initial zone of agreement should itself be identified. Who convenes the first participants? How diverse must they be? How are excluded voices detected? How are competing purposes for technological development surfaced and compared? Who gets to challenge purposes that have come to be treated as self-evidently desirable? How are conflicts of interest exposed? How is apparent consensus distinguished from genuine convergence?

These questions cannot be answered merely by announcing another grand institution from above. They are part of the bootstrap problem itself.

We do not need universal agreement on every ethical controversy before we can improve collective decision-making. We need a sufficiently robust initial zone of agreement and practice – a set of principles that diverse people can endorse for different reasons, combined with procedures and behaviours that can begin to earn warranted trust. From that base, we can:

- Build institutions and tools that deserve a limited degree of trust
- Use these institutions and tools to scrutinise more powerful systems
- Where justified by evidence, expand the circle of warranted trust.

The aim is not ever-expanding *blind* trust. It is ever-expanding *warranted, bounded, contestable, and revisable trust*.

That is the growing spiral of technoprogressive trust.

Endnote

Many of the questions raised in this essay will be explored further at *The Technoprogressive Opportunity: The Future of Ethics and Emerging Technologies*, a conference taking place in London on 19-20 September 2026.

For further information, see: opportunity.londonfuturists.com.