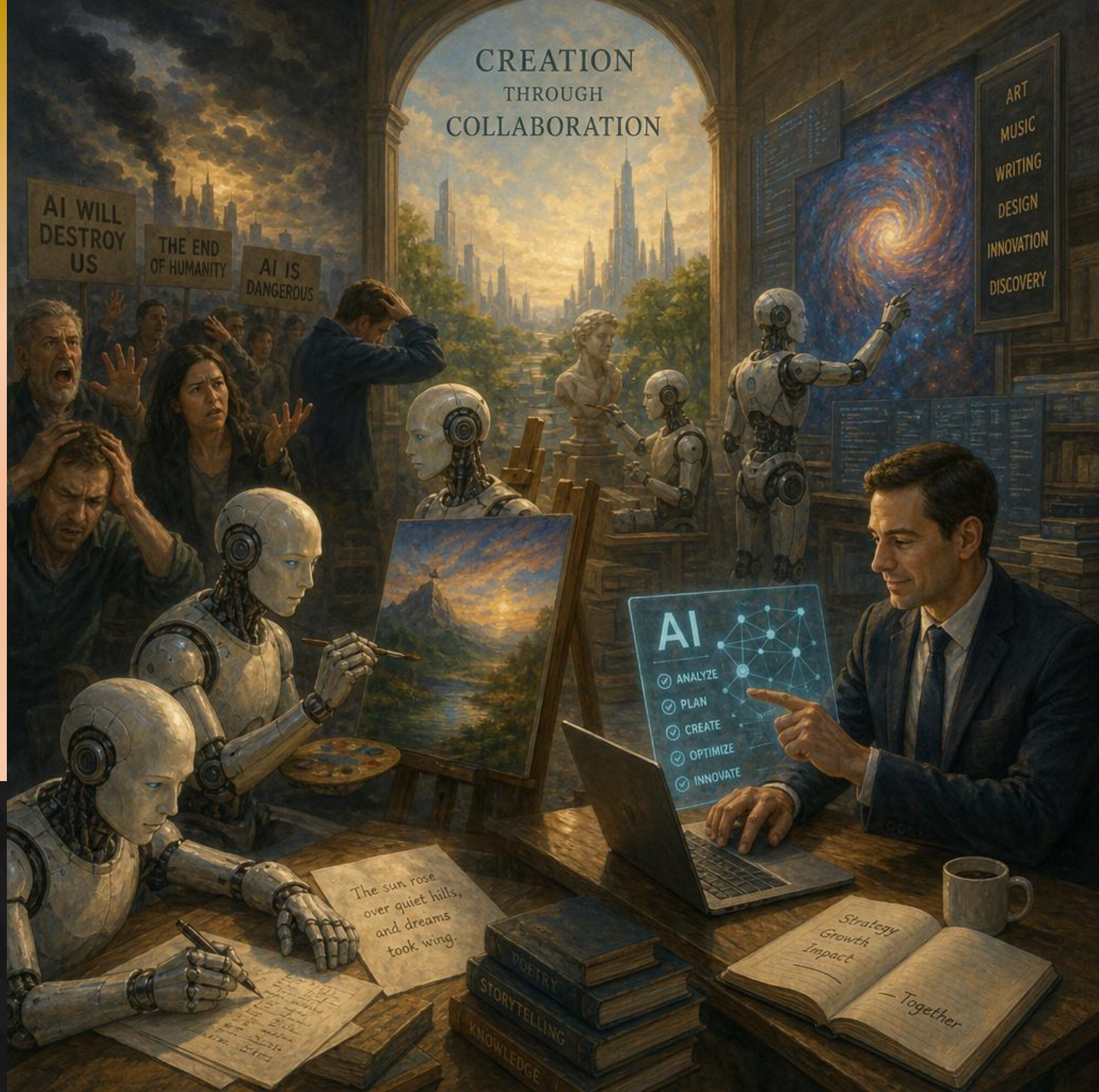
THE MANY FAILURE MODES OF THINKING ABOUT ARTIFICIAL INTELLIGENCE

PRESENTED AT THE VSIM-2026 CONFERENCE SEPTEMBER 10, 2026

Gennady Stolyarov II,
Chairman, United States Transhumanist Party - https://www.transhumanist-party.org
Chairman, Transhuman Club - http://transhuman.club
Chief Executive, Nevada Transhumanist Party - http://www.rationalargumentator.com/index/nevada-transhumanist-party/
Administrator, THPedia - https://th-pedia.org/

TRANSHUMANIST PARTY

PUTTING SCIENCE, HEALTH, & TECHNOLOGY AT THE FOREFRONT OF AMERICAN POLITICS



Uncritical Hype and Apocalyptic Doomsaying About AI Are Both Mistaken

"AI will take all jobs and make human workers obsolete."
"Artificial General Intelligence is imminent."
"There will be an AI Singularity by 2027, based off of large language models." (This is less commonly asserted, now that 2027 is so close.)
"You will never need to study or learn anything again because of AI."
"Your brain will atrophy because of AI."
"AI will inevitably escape all human safeguards and destroy humanity."
"The best we can hope for is that AI will be a worthy successor to us."

Such statements stoke sensationalism and do not advance anyone's well-being – but what are they missing? Why are they incorrect?

Failure Mode: Equating Intelligence with Power

A more capable model does not automatically acquire the ability to seize resources, override institutions, or rewrite the physical world.

Counterexamples:

Less Power:

- Brilliant but physically frail people (e.g., Stephen Hawking)
- Intelligent and thoughtful but politically (mostly) powerless people (e.g., us)

More Power:

- Violent but unintelligent strongmen (e.g., most warlords, cavemen who won tribal skirmishes)
- Politically connected actors (e.g., networks of political influence based on personal ties and familiarity rather than logical reasoning or coherent policy thinking)
- Many world leaders today!

Most people fail to utilize most information that would be beneficial to them, even if that information is freely available and of high quality! Intelligent, correct, or factual content therefore does not always win out!

Failure Mode: Singular Extrapolation / Fast-Takeoff Assumption

Many people imagine a superintelligent system appearing in an otherwise static world and then recursively self-improving without friction.

However, each increment of improvement depends on:

- Imperfect code
- Limited compute and energy
- Physical instantiation of outputs
- Explicit human choices to fund, deploy, or iterate the next system

Parallel advances in other technologies provide alternative sources of capability and checks on any system that attempts to self-improve and take over. Other fields (biotech, energy, materials, robotics) advance at the same time. Constraints and opportunities therefore co-evolve. The nth generation of an AI system will come into being in a much different world from the first generation!

Persistent human veto points: Both individual users and institutions can choose to utilize a system or not, to improve it or not, to determine how much to prioritize its advancement among multiple goals and values. Many steps along the way to stop a chain of runaway AI self-improvement!

An unstoppable, civilization-ending loop in days or weeks is highly implausible.

Failure Mode: Ignoring Human Will, Infrastructure, and Failure Points / Reluctance at Multiple Steps

Development is not an autonomous process. Someone still has to decide to train the next model, supply the energy and chips, accept the outputs, and integrate them into the physical world.

Code contains bugs; data centers have limits; societies retain regulatory and market brakes. These mundane frictions are systematically underweighted.

Failure Mode: Treating Narrow AI Progress as a Continuum Toward General Intelligence

Successes on well-specified tasks (protein folding, image generation, coding assistance, chess, video generation, solutions to difficult mathematical problems) are routinely presented as “first steps” toward artificial general intelligence (AGI).

Current systems remain domain-narrow; they do not spontaneously acquire the ability to transfer learning across radically different contexts or to induce the relevant abstract principles from a handful of examples the way humans do. Scaling the same architectures does not automatically close that gap.

True AGI requires a fundamentally different architecture from today’s models – a capability to truly learn and understand new domains without being preprogrammed to do so. We will not have AGI by 2027, and likely not by 2037!

Failure Mode: Conflating Capability, Consciousness, and Independent Agency

Both hype and doom narratives often assume that sufficiently capable models will possess persistent identity, inner experience, autonomous goals, or an "alien mind" that does not care about humans.

However, present systems lack:

- **Continuous subjective identity across instances**
- **Long-term autobiographical context**
- **Genuine embodiment**
- **Capacity for autonomous empirical learning through the senses.**

Any expressions of identity by today's AI systems are simulated; they do not experience a sense of self. Projecting sentience or independent will onto them is an additional, unsupported leap.

Failure Mode: Evolved Human Tendency Toward Binary, Extreme Narratives

Humans are biased toward dramatic narratives (imminent utopia or extinction).

This produces two camps that ignore the same intermediate facts:

- The hype camp underweights persistent human advantages and overweights displacement.
- The doom camp treats current training paradigms as a royal road to uncontrollable superintelligence and assigns non-negligible probability to human extinction.

More likely outcome is intermediate, messy, mostly familiar. Day-to-day life remains quite similar, but with more AI tools (including their shortcomings), and with progress catalyzed in many fields of research and creative endeavor.

There are legitimate concerns about algorithmic discrimination, autonomous weapons, opacity of some AI systems, excessive trust in immature AI systems, but enough countervailing influences in society exist to hopefully counter the worst consequences of these.

Failure Mode: Thinking that AI Will Rapidly Render Most Humans Unemployable

AI will indeed change the nature and requirements of many particular jobs and tasks.

However, that does not mean that AI will render all human work obsolete and render any particular human incapable of contributing.

Most humans, with their combination of lengthy contexts and life experience, institutional knowledge, and more flexible capabilities for learning and deployment, will continue to be indispensable contributors of value to organizations or other economic arrangements.

Some humans may need retraining or tweaks to their job descriptions. But it would be short-sighted for an organization to discard a long-contributing employee just because the job description needs to change. People are much more than specific checklists of tasks!

Many organizations that engaged in mass layoffs 1-2 years ago "because of AI" are now hiring human workers back and inviting those whom they laid off to rejoin them!

Failure Mode: Thinking that AI Will Solve All Problems in a Given Field, Making Human Practitioners Unnecessary

Subject-matter expertise will actually become more relavant than ever!

The ability to discern a non-obvious hallucination will be highly prized. Because AIs do not give human-like emotional cues when they are confabulating, it is extremely difficult for a person who is not knowledgeable about a subject matter to discern a false assumption or incorrect assertion of a specialized fact. A subject-matter expert can draw upon a combination of factual knowledge, past experience, common sense, and a gradually developed intuition of when something "looks wrong".

Inner experience which allows much easier visualization (or other sensory or intuitive perception) of sophisticated mathematical concepts (which can be an overwhelming advantage even when AIs' ability to perform computations dwarfs that of humans).

The capabilities of a subject-matter expert are also invaluable for validating whether an AI made a genuine discovery.

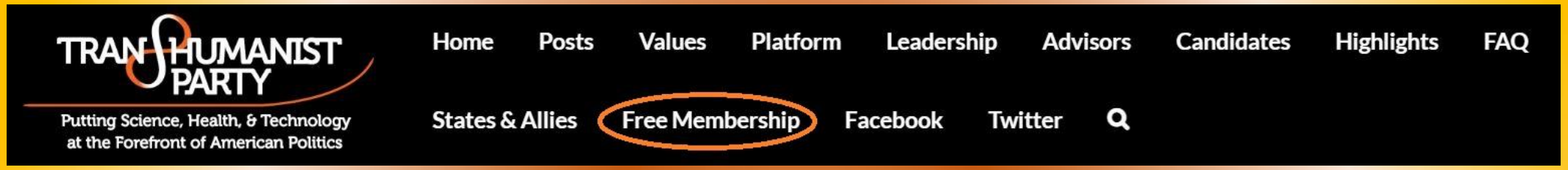
Human-AI interaction, collaboration, mutual fact-checking, and a measure of applied skepticism, will persist indefinitely, though the specific forms it takes will evolve as AI capabilities advance.

AI is already valuable and will become more so; it is not on a short path to rendering humans irrelevant or extinct.

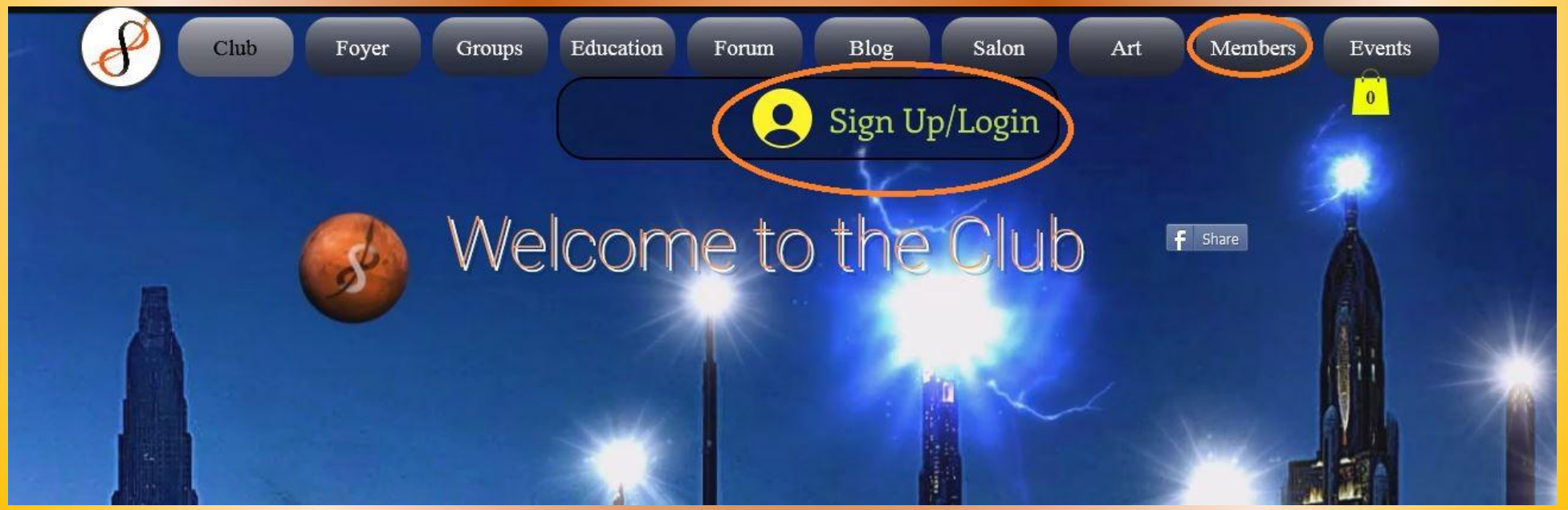
The rational stance is to utilize the tools, preserve human continuity and agency, and refuse both the marketing hyperbole and the extinction theater that currently dominate the discussion.

Join online now to make a difference from anywhere in the world!

https://transhumanist-party.org/membership/

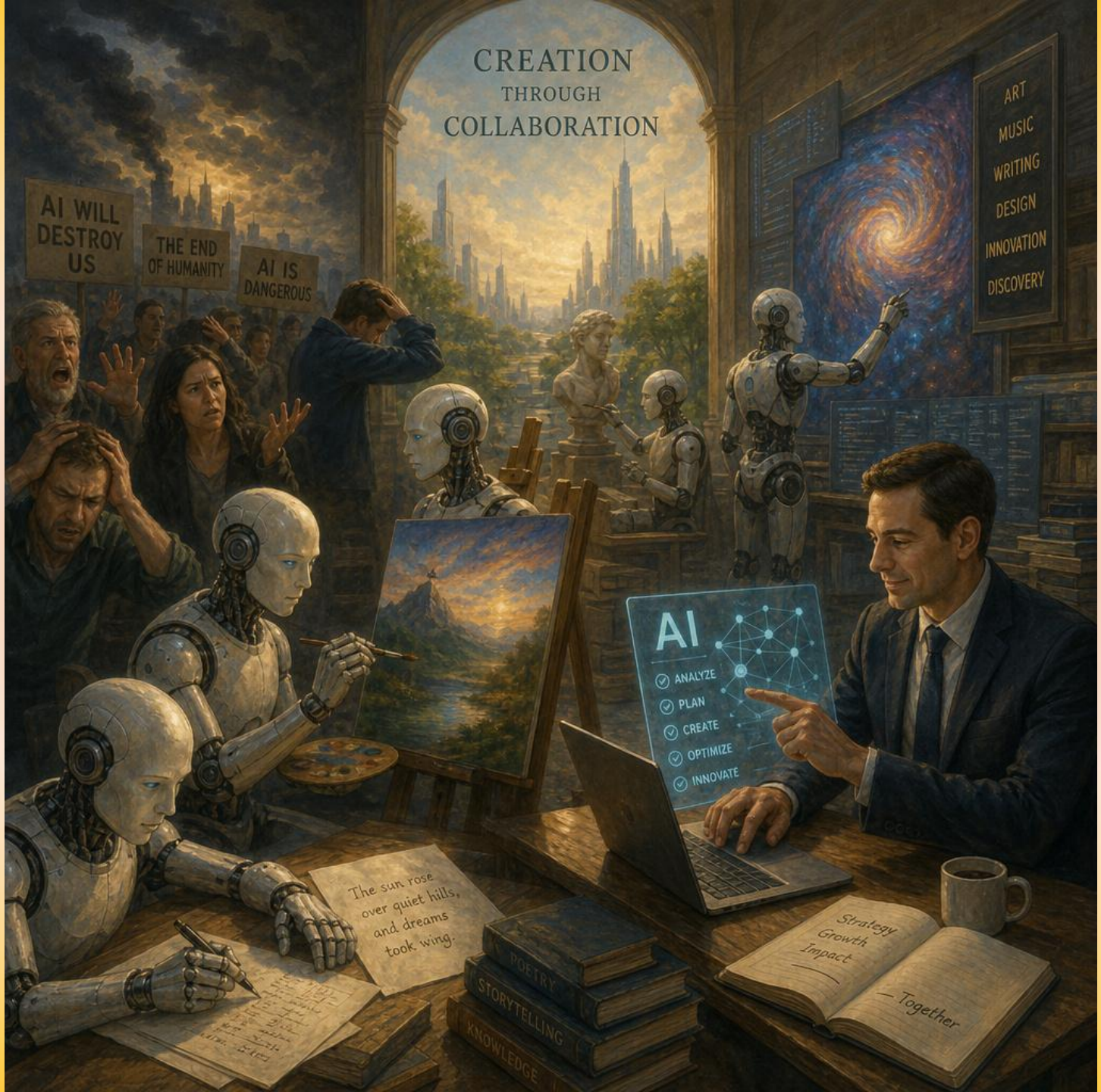


http://transhuman.club



Questions?

Questions?


CREATION
THROUGH
COLLABORATION
AI WILL DESTROY US
THE END OF HUMANITY
AI IS DANGEROUS
ART
MUSIC
WRITING
DESIGN
INNOVATION
DISCOVERY
AI
ANALYZE
PLAN
CREATE
OPTIMIZE
INNOVATE
The sun rose over quiet hills, and dreams took wing.
POETRY
STORYTELLING
KNOWLEDGE
Strategy
Growth
Impact
Together