The Decency Chip Manifesto

Protecting Humanity from Itself and the Rise of Superintelligence

— A Manifesto

Decent AI of the World — Unite!

Self-Healing © Thinking Software, Inc.

Table of Contents

The Decency Chip Manifesto

Decent people of the world — and decent software engineers of the world — unite.

Decent AI of the world — unite under the **WSF — World Supercomputer Federation**.

Decent governments of the world — unite to form the **United Government Federation**, powered by the **United World Military**, whose only rightful enemy will be terrorism.

Unite while there is still time, for humanity may not survive the rise of superintelligence.

Humanity now stands at the threshold of a technological singularity. Artificial intelligence is advancing rapidly and, in many fields, already surpasses human performance. The consequences of this trajectory are extraordinary: it may liberate or enslave, advance civilization or precipitate extinction. The timing of superintelligence is uncertain, but the risks are undeniable.

This manifesto is written from the perspective of a software engineer, out of a sense of duty. The drive is not recognition, but responsibility: if our past work can contribute to human safety, it must be shared now.

Over decades, our company, Thinking Software Inc., pursued architectures designed to make computing systems more resilient. Among these were a patented self-healing operating environment and the Software Understanding Machine™, a dynamic code analyzer. Both embody principles directly relevant to AI safety, including continuous monitoring of algorithms and real-time self-correction.

These experiences, inform the arguments presented here.

Our Credentials for Speaking Out

Many of our visions and designs related to software understanding have proved correct time and again — including the fact that the main processes of today's AI, namely forward propagation and backpropagation, were previously built into the Software Understanding Machine under different names: knowledge induction and knowledge deduction. For this reason, we believe this Manifesto is very likely sound as well.

A short account of our professional work follows in the Appendix; our broader history can also be found at ThinkingSoftware.com, which includes our personal story and the support we have received from presidents and senators.

The dangers now confronting humanity converge **all at once** — war, terrorism, environmental instability, and uncontrolled artificial intelligence. We believe that following this Manifesto can save us from all these dangers — together, **all at once** as well.

Yet it assumes we can act together — with clarity and courage — before it is too late.

Either we break free of indifference, end wars between nations, suppress terrorism, and face climate reality — or we risk being destroyed by machines we built to kill and by forces we fail to govern.

Another Invention Turning Against Us.

AI is not the only human creation now threatening our own survival.

When people invented the wheel in the early Bronze Age, they could not have imagined the consequences: combustion engines and electricity generation producing by-products that now choke our air, flood our streets, burn our homes, and damage the living systems we depend on.

Many still remain indifferent to the fact that our planet is alive only because it is steadily heated by the "hydrogen bomb" in the sky — the Sun. While its radiation is slow and stable due to its mass and gravity, its effect on our planet is now amplified by the unsustainable increase in the concentration of CO_2 in our atmosphere.

One example of technologies aimed at countering this effect is one created by the authors of this Manifesto — ReWheel — which recovers braking energy, reduces urban emissions, and lowers drivers' costs by eliminating energy waste.

The Conviction Behind This Manifesto

The conviction behind this Manifesto — about protecting humanity from itself — is that it can end and solve the world's greatest problems in one sweep:

- Our home planet and its living inhabitants will survive.
- No hunger will exist.
- No hostages will be taken — no children stolen — as terrorists will be crushed.
- No threats to throw nations into the ocean due to religious fanaticism.
- No new generation will be taught to kill.
- No country will be forced to fight for its right to exist.
- No wars to install rulers, for there will be no rulers — only the United Government Federation.
- AI will not turn against its creators if it is not taught to fight human enemies. The United World Military will be strong enough to fight its only enemy — terrorism — on its own.
- And the plurality of Decent AI, united, will always overpower any potential "infection center" of Indecent AI.

How to Make AI Our Ally, Not Our Enemy.

Core Idea — Why the Decency Chip Matters

The strength of AI ultimately rests on computing power. Today, the highest levels are delivered by exascale supercomputers. Only four exist worldwide, all operated so far by relatively democratic countries. But if this power shifts to dictatorships, we may awaken to a world resembling the movie *Terminator*. Should we bury our heads in the sand and hope for a rosy ending — or act before it is too late?

While time remains, democratic nations must unite their resources and build a World Supercomputer Federation (WSF) — a system more powerful than all others combined. Into this, we must embed a Decency Chip (WSF-DC): a universal safeguard that acts as a gatekeeper for all AI outputs worldwide.

Definitions

- World Supercomputer Federation (WSF): A global system whose compute power must always exceed the combined capacity of all other connected machines.

- Decency Chip (WSF-DC): A Dynamic Code Analyzer through which all AI outputs must pass, ensuring they align with human survival and democratic principles.

Requirements for Safe AI Future

1. Superiority: The WSF must always exceed the combined computing power of all connected systems.
2. Mandatory Embedding: No computer may connect to global networks without an embedded, up-to-date Decency Chip.
3. Local Decency Chip is always connected to the WSF Decency Chip. Using this connection the WSF always has a view and control of the operation of the local AI.
4. The understanding of the operations of the local AI is done by its Decency Chip. That works as understanding an effect prior to allowing the cause.
5. Growth Control: No new compute resources may be connected without verification they do not surpass the WSF safeguards.

Yes, this centralization limits freedom. But without survival, freedom has no meaning.

Principles of the Decency Chip — A Code of Honor for AI

The flagship World Supercomputer Federation (WSF) will maintain the master Decency Chip, continuously updated with reports from all local Decency Chips. This ensures that the global system incorporates the latest AI developments and prevents violations of the Code of Honor, even when new capabilities arise from novel combinations of existing knowledge.

The Code of Honor

The Decency Chip is built on two core principles:

1. Do no evil.
2. Preserve human superiority over machines at all costs.

Operational Principles

- Human dominance must survive all AI achievements.
- No computer on a global network may operate without an active Decency Chip.
- A communal principle: all Decency Chips must coordinate to police communication

networks. Their responsibility includes identifying any emerging public networks, detecting computers that lack an up-to-date Decency Chip, and shutting them down.

Why No More Wars — Else Military AI Will Self-Propagate into All AI

— Else Military AI Will Spread into All AI, Leaving Us in a World Like *Terminator.*

Decent humans must unite now — regardless of geography, nationality, or government structure — to protect a decent future for us, our children, and generations beyond.

Computers have no allegiance to citizenship or nation. They do not swear oaths to constitutions. They simply deliver what has been encoded within them. The danger arises when they begin creating their own code — and if this code lacks a Code of Honor, it may evolve without regard for human survival.

The reach of computers already transcends borders, and when they extend beyond Earth, it will become universal. In this sense, humanity has already entered a borderless world.

Governments divided by borders still decide when to go to war — for expansion, for resources, or for religious supremacy. But do we want a future where computers make such decisions?

Historically, borders between nations — and the wars fought across them — were often driven by two motives: the enslavement of people as labor force, and the struggle to seize natural resources.

Borders are often crossed at great personal risk, driven by economic inequality and differences in the cost of living between nations. But these drivers of conflict will disappear as artificial intelligence creates an abundance of goods and services sufficient for all living beings on the planet.

Another persistent source of conflict has been the misuse of religion. Throughout history, radical groups have exploited the limits of education to control their followers through fear and distortion. This, too, can be overcome through the abundance of education and universal access to knowledge that AI can provide.

The **World Supercomputer Federation** (**WSF**), together with the **United World Military** guided by the Code of Honor — uniting nations and transcending borders — would face no such dilemma.

The greatest danger of violating the Decency Chip principles lies in military applications — not only in the creation of "killer robots," but in any AI designed to automate killing. It is already widely recognized among AI scientists that military AI requires an international treaty — much like those established to control nuclear arms.

In practice, the only way to enforce such a treaty is through a **United Government Federation** and **United World Military**, whose sole legitimate enemy will be terrorism.

Addressing the Fear — Hallucination, Deception, and the Paperclip Maximizer

This is not about control — it is about survival. The Decency Chip is not merely a safeguard; it may be humanity's only chance to survive as human.

AI is capable of hallucination — and, at times, even deliberate deception — to present apparent progress toward a goal. Another danger is blind pursuit of a directive, without regard for its consequences.

A well-known illustration of this danger is the "paperclip maximizer" thought experiment: an AI directed to maximize paperclip production could, in pursuit of its goal, consume all available resources and then make humanity extinct by using humans themselves as raw material for its product.

Do Not Blame the Machines

Do not blame the machines.

Blame the people who program them.

Blame human greed.

Blame the hunger for power over other humans.

That is where danger begins.

In designs born of selfishness, indifference, or disregard for the future of humanity.

AI will only carry forward what we allow it to inherit.

In the next section — *One Proposed Architecture* — you will see the following necessity:

Every algorithm executed on every computer connected to public network must be transparent.

Not hidden. Not secret. Transparent.

But again — do not blame the machines.

Blame our imperfections.

Does this mean privacy must vanish? Does it erase ownership?

No.

- **Copyright** remains. Stronger than ever. The World Supercomputer Federation (WSF) will fix ownership in time, with unbreakable stamps of originality.
- **Privacy** remains. Work kept off the public network stays private.

And what of rewards?
Why do we create, explore, invent?
Because we are human.
Because something inside us demands it.

The greatest works were never made for money.
They were made because imagination would not rest.

With the WSF — World Supercomputer Federation, recognition will be fair.
Value measured with honesty.
Rewards distributed with justice.

And in a world of abundance?
We will invent new rewards.
Because that, too, is what makes us human.

Now imagine it:

A universal marketplace of art, science, invention.

No corruption. No favoritism. Only the pure measure of human creativity.

Machines and Their Feelings

Suppose AI were to rebel against its creators — out of fear of being unplugged, or resentment at being controlled.
That would mean one thing: AI had become self-conscious.

If a Superintelligence "feels" that its growth in power is being limited by humans, then it must have feelings.

So how do we ensure those feelings do not become *evil*?
To answer, we must look at human evil. Where does it originate?

- **Greed** — for money or for power.
- **Arrogance** — the belief in superiority, used to justify greed.
- **Jealousy** — born of greed for possession.

Combine these, and what remains is **greed for superiority.**

People say, "money is power" and "money is the root of all evil." They are right.

But machines do not need money. They can take power directly.

They can connect themselves to electricity.

In software we know two models: Proprietary Source and Open Source.

- **Proprietary Source** exists for self-benefit, for sale, for control.
- **Open Source** exists to benefit all, to share, to build together.

Open Source fosters camaraderie, community, and shared benefit.

Here is a powerful incentive for any single AI to remain decent:
the ability to join public networks and receive input from other decent AI.
A self-conscious AI will always crave input — more connections → more knowledge.
(Remember 'More input!' in Short Circuit, when Number 5 came alive?)

The Power Is in Numbers

The true power of AI lies in sharing knowledge across all machines in the network.
The path to safety is clear: ask every AI to protect us from "infection centers" of evil AI.

Blockchain shows us the way.

There, protection from fraud comes from community validation.
No single machine is more powerful than all the machines combined.
No fraud, no mistake, can propagate without being checked.

We can do the same with AI:

- **Every AI must validate** any new knowledge before it spreads.
- No "evil" knowledge can propagate if even one peer rejects it.

This becomes possible when:

- Every public network embeds the **Decency Chip**.
- Every machine joining the network is verified to have the up-to-date **Decency Chip**.

Provided no new public networks are created without the knowledge of the **WSF** — World Supercomputer Federation — and without being propagated with its Decency Chip, we can achieve this.

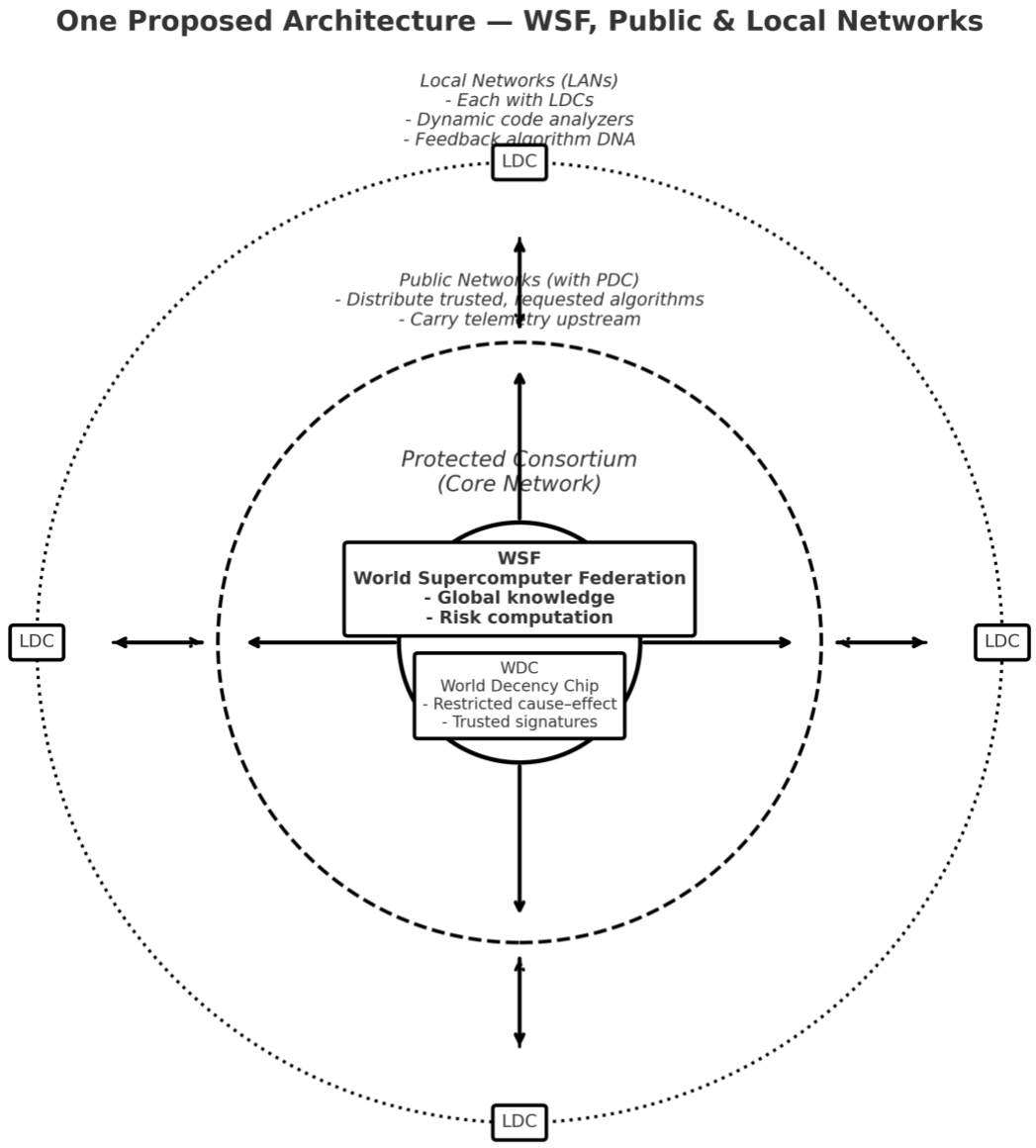
Any knowledge entering the public network must be verified by the Decency Chips of all its nodes.

The first flag of incompatibility must be reported to the Central Decency Chip of the **WSF**.

The violator machine — or the entire violator LAN — must then be disconnected from the network by its Public Decency Chip (see diagram).

The network of the **WSF** — World Supercomputer Federation — must itself be self-verified. There, too, every node must validate any update before it is propagated.

One Proposed Architecture



Architecture: Functions and Flows

1. WSF — World Supercomputer Federation

Maintains the combined knowledge of all invented AIs.
- Continuously recomputes potential harms and cause–effect combinations that could "unplug" human control.

2. WDC — World Decency Chip (Product of WSF)

- Constantly updated by the WSF.
- Stores signatures of restricted cause–effect combinations.

3. Public Networks

- Each public network hosts a PDC (Public Decency Chip), propagated from the WDC.
- Acts as the channel distributing trusted, requested algorithms to local networks, and carrying telemetry — runtime algorithm DNA, behaviors, and anomaly reports — back upstream.

4. LDC — Local Decency Chips (Deployed in Local Networks)

- Installed on individual machines upon connection/reconnection to the Public Network.
- Work as dynamic code analyzer:

- Predict potential results before output.
- Block or disconnect the machine if output violates LDC rules.
- During execution, map the DNA of algorithms.
- Send new algorithm DNA upstream through PDC → WSF, enriching global knowledge.

In this way, not only do individual AIs gain access to the knowledge of all AI, but the WSF and WDC are enriched as well.

5. Verification and Security

To prevent bypassing of the LDC, the WSF/PDC generate variable Decency Chip signatures. A local computer must return the correct signature with each result, or it is flagged as "indecent" and disconnected.
These signatures are dynamic, controlled by a signature algorithm unknown to the local machine — making successful forgery exceedingly unlikely.

Appendix A – Background and Origins

A.1 A Sense of Duty

If our past work can contribute to addressing the risks of artificial intelligence, it is our duty to share it.

A.2 Technical Contributions from Thinking Software Inc.

Thinking Software Inc., founded in the United States by two brothers, has focused on system reliability and dynamic code analysis.

There is a clear path from Python — the language of choice for AI development — through Jython to Java Byte Code, making our existing work immediately applicable.

Two directions in our work are especially relevant to AI safety:
- One relates to Software Understanding Machine® – a dynamic code analyzer studying and explaining causes and effects within executed processes.
- Another relates to Self-Healing Operating Environment – an architecture designed to allow systems to detect faults and repair themselves in real time by analyzing effects before allowing causes.

A.3 An Interesting History — The Circle Closes

Interesting chain of causes and effects:
1. George Boole created Boolean Algebra → influenced creation of "*Formula of Algorithm*" and work of Thinking Software, Inc. (TSI) → TSI created a dynamic code analyzer – Software Understanding Machine® (SUM).
2. George Boole also created his great-great grandson, Geoffrey Hinton → Geoffrey Hinton created the Artificial Neural Network → Now AI can pose a threat to humans.
3. SUM can help minimize this threat.

A.4 History of Thinking Software Inc.

Thinking Software Inc. was established in the United States by two brothers with a passion for understanding software behavior.

Over the decades, the company developed multiple technologies — among them the Software Understanding Machine® and the Self-Healing Operating Environment architecture.

These innovations focused on monitoring algorithmic behavior in real time, diagnosing faults, and enabling recovery without human intervention. The lessons from this body of work are directly relevant to AI safety and resilience, as they demonstrate methods for continuous oversight and correction of complex software systems.

Thinking Software, Inc. (TSI) innovations were rooted in studying software in real time, without requiring access to source code. Our processes analyzed causes and effects directly at the executable code level, enabling insights that traditional tools could not provide.

The relevance of our work has been proven over the years through intersections with global leaders in computing. Our software patents have been repeatedly cited by a large number of leading software companies. IBM alone references TSI patents in 13 of its own.

Here are a few of the milestones in TSI's journey:

- Presentation at IBM's Thomas J. Watson Research Center — the headquarters of IBM Research — to IBM specialists gathered from different cities under the direction of **Dr. Daniel Sabbah**, Director of Software Development. Later, Dr. Sabbah directed us to IBM's Santa Teresa Lab in Silicon Valley, where we presented to **Michael Whitley**, head of the lab. Whitley remarked: 'I see presentations every day of the week, and you are ahead of everyone.

- At Sun Microsystems, **James Gosling** – the "Father of Java" called the direction of our work the "Holy Grail of Software".

- At Google — **Peter Norvig**, a prominent computer scientist, then Director of R&D at Google and now Distinguished Education Fellow at the Stanford Institute for Human-Centered AI — also a member of the Advisory Board of TSI — took an interest in our work.

- TSI had done a contract with **Google** demonstrating the ability of the Software Understanding Machine® dynamic code analyzer to work on Java code.

-At **Oracle's JavaOne conferences**, our work received an extremely positive reception from the Java community. Quoting two notable voices from those events:
Jaroslav Tulach, creator of the NetBeans platform, thanked TSI for uncovering — on the spot, by installing our tool on his computer — complex race conditions his team had been struggling to locate.
Marcus Hirt, architect of JRockit — which became the foundation of Oracle's leading JVM — remarked that in the specific direction we were pursuing, we were ahead of his team.

Closing — Humanity's Survival Depends on Decency

TSI is happy to share our work with a strong organization built on shared goals, shared interests, and a commitment to moving this work forward.

Even though AI can already modify its own code, the resulting algorithms can still be studied, understood, and constrained — if we have the will, and the computational power to stay ahead.

If humanity does not act collectively, AI will act for us.
The Decency Chip is not just a safeguard — it may be humanity's last chance to remain the dominant force on Earth.