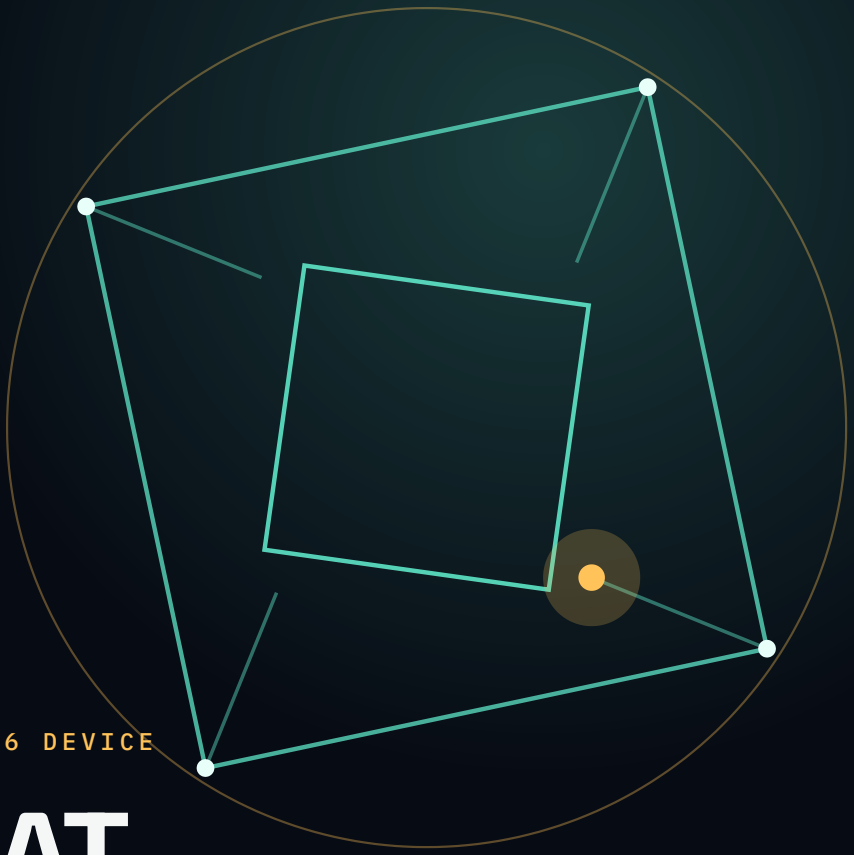
ai021

WHITE PAPER NO. 1



FROM A 2022 THEORY TO A 2026 DEVICE

From HARAT to NanoSI

Human-centered, verifiable intelligence at the edge: how a theory of human–AI co-thinking became private, offline, source-linked AI.

"The human mind and artificial intelligence can achieve better results when collaborating with a shared purpose." Khan, 2022

Zafrullah Khan, Ed.D.
Founder & CEO, ai021.ai, Inc.

VERSION 1.0 · NOVEMBER 2026
WWW.AI021.AI/RESEARCH

ABOUT THIS PAPER

From HARAT to NanoSI

Human-centered, verifiable intelligence at the edge: how a theory of human–AI co-thinking became private, offline, source-linked AI.

Author. Zafrullah Khan, Ed.D., Founder & CEO, ai021.ai, Inc.

Series. ai021 White Paper No. 1 · Version 1.0 · November 2026

Audience. Investors and venture partners, leaders of organizations that cannot afford unreliable or leaky AI, educators, and anyone who wants AI that keeps people in charge.

Origin. This paper builds on *Beyond Critical Thinking and Creativity: Human Adaptive Reflexive Agile Thinking (HARAT)*, a peer-reviewed workshop the author presented at the 20th Annual Hawaii International Conference on Education, Waikoloa, Hawaii, January 3–6, 2022. The original workshop materials are available unchanged at www.ai021.ai/research. Where this paper quotes the 2022 work, the words are the original; where it interprets them, it says so.

CONTENTS

Inside this paper

– Executive summary

01 Origins: Waikoloa, January 2022

02 The problem, four years later

03 HARAT revisited: adaptive, reflexive, agile

04 From theory to architecture: NanoSI

05 Intelligence at the edge

06 Proof points: the first device and the Sticky suite

07 The learning hierarchy, applied to machines

08 Where it matters

09 Trust by design

10 Roadmap: from one device to an AI operating system

11 What this means

– Conclusion · About the author · About ai021 · References

Important notice. This white paper describes research, products in development and plans. Statements about future plans are aspirations, not promises or an offer of securities; actual results may differ. Nothing in this paper is an offer to sell, or a solicitation of an offer to buy, any securities. Capacity figures marked “projected” have not yet been confirmed by load testing. Product images are illustrations, not screenshots; real interfaces are shown in live demonstrations. Use-case scenarios are examples, not customer case studies. © 2026 ai021.ai, Inc. All rights reserved. “ai021”, “NanoSI”, “StickyWork”, “StickyMesh” and “StickyStudio” are names used by ai021.ai, Inc.

EXECUTIVE SUMMARY

The idea came first. The device came next.

In January 2022, nearly a year before ChatGPT made generative AI a household word, an educator stood before a room of teachers and researchers in Hawaii and argued that the future of artificial intelligence would not be decided by how powerful machines became, but by how well people and machines learned to think together.

The framework was called **Human Adaptive Reflexive Agile Thinking (HARAT)**. It made a handful of claims that were unusual for an education conference at the time. Human intelligence and artificial intelligence would form recursive feedback loops, each improving the other. AI would be limited by the information it is given, while critical thinking would remain stubbornly human. And machine intelligence would become *"embodied, embedded, distributed, situated, 'ambient' and pervasive."*

Four years later, AI capability has exploded and trust has not kept pace. The dominant models are enormous, live in distant data centers, need a constant internet connection, absorb whatever their users type into them, and answer fluently whether or not they are right. For a school with unreliable Wi-Fi, an offshore platform on a satellite link, a law firm guarding client files or a government office that is not allowed online at all, that model of AI is somewhere between inconvenient and unusable.

ai021.ai, Inc. was founded in 2026 to build the alternative HARAT pointed toward. Its approach, **NanoSI** (nano structured intelligence), builds **small language models from scratch**, **fine-tunes** them for a defined job, and runs them **on edge devices, offline**, answering from structured, source-linked knowledge, so that every answer can be traced back and checked and an unsupported answer stays unsupported.

This is not a slide. The first NanoSI device, a portable USB unit, is built and ready to demonstrate. Three products run on it: **StickyWork** (productivity tools), **StickyMesh** (offline collaboration and conferencing) and **StickyStudio** (small language model inferencing). Over a local router it is designed for 300–500 simultaneous users (projected, to be confirmed in load testing).

KEY TAKEAWAYS

1 **A thesis that predates the boom.** HARAT (2022) described human–AI feedback loops, the limits of machine knowledge and ambient, embedded intelligence before generative AI went mainstream.

2 **The market has moved toward it.** Capable small open models, documented local serving tools and evidence that smaller, well-trained models can win on defined tasks make edge AI practical now.

3 **Offline is a feature, not a fallback.** The device works as a backup when the internet fails and as a permanent system where the internet is absent, unreliable or not allowed.

4 **Privacy and IP protection by architecture.** Data, prompts and models never have to leave a device the customer physically controls.

5 **Built and demo-ready,** with a roadmap from one device to a family of edge devices and, next, ai021's own AI-native operating system (AIOS).

AI021 · WHITE PAPER NO. 1



SECTIONS 01 – 03

PART I

The idea came first.

Where HARAT came from, what has changed since 2022, and why the problem is now urgent.

01

ORIGINS

Waikoloa, January 2022

Some companies start with a product and look for a philosophy. ai021 started with a philosophy, tested in front of peers, and spent four years turning it into hardware.

Zafrullah Khan spent three decades in education before founding ai021: teaching in Durban, South Africa, and at a tribal school in South Dakota; working as a school psychological examiner; leading special education for a school district; directing research for the Tennessee Department of Education; and preparing new teachers as an associate professor at Middle Tennessee State University from 2005 to 2022. Assessment work leaves a lasting habit: a single score rarely tells the whole story, and a person's judgment belongs in the picture.

That habit shaped the workshop he presented, after peer review, at the 20th Annual Hawaii International Conference on Education. Its title was long and precise: *Beyond Critical Thinking and Creativity: Human Adaptive Reflexive Agile Thinking (HARAT) using different combinatoric neural network architectures for supporting human decision optimization when working with artificial intelligence (AI) in education.*

> *"Critical thinking will never be fully automated because critical thinking draws on the unique perspectives of the one being critical."*
>
> KHAN, HARAT WORKSHOP ABSTRACT, 2022

The workshop asked educators to treat AI as neither a threat nor a toy, but as a partner that demands discipline: people and machines achieving more together, as long as human judgment, ethics and perspective stay at the center.



Figure 1. From theory to device. The grey milestone is an industry event for context, not an ai021 milestone.

Three ideas that aged well

FEEDBACK LOOPS

Two learners, one loop

HARAT described "recursive iterative feedback loops generated by both learning systems," human and machine, each sharpening the other.

LIMITS OF MACHINES

Only as good as its sources

"AI is limited to the data and information that is programmed." Today we would say: a model should answer from its sources, or admit it can't.

AMBIENT AI

Intelligence moves into things

Machine intelligence would become "embodied, embedded, distributed, situated, 'ambient' and pervasive." That is edge AI.

In 2022 these were questions for educators. ai021's answer is architectural: AI small enough to own, honest enough to check, and close enough to its users to keep working when the network does not.

02

THE PROBLEM

The problem, four years later: capability without control

AI became dramatically more capable after 2022. It did not become proportionally more trustworthy, private or available where people actually work.

1. Scale is expensive, and often unnecessary

The industry's default answer to every problem has been a bigger general model running in someone else's data center. Yet much valuable work is narrow and repetitive: classifying requests, answering from a manual, summarizing a meeting, drafting from a template. Research on compute-optimal training (Hoffmann et al., 2022) showed that a smaller model trained on more data could outperform much larger models of its time. Our own published test points the same way at small scale (Section 6). Paying for a giant generalist to do a specialist's job is a cost structure, not a law of nature.

2. Black boxes cannot be audited

When a model cannot show where an answer came from, compliance review becomes guesswork. Fluent text is not evidence. In regulated work (finance, healthcare, law, safety, government), an answer without provenance is a liability no matter how confident it sounds.

3. The connectivity assumption

Cloud AI quietly assumes a fast, constant internet connection. Offshore platforms, ships, mines, farms, field research stations and many rural schools do not have one. Offices lose theirs during outages. Some facilities forbid it. For these settings, AI that stops when the connection drops is not a product; it is a dependency.

4. Data and intellectual property leave the room

Every prompt sent to a cloud service carries something: a contract clause, a design, a student record, source code, a strategy. Whether that information is retained or used for training depends on a provider's terms, which customers do not control and which can change. For organizations whose value *is* their know-how, that is a structural risk.

5. Cognitive outsourcing

The quietest problem is the one HARAT worried about most. Tools that answer instantly and confidently, without showing their sources, train people to stop checking. Over time, human judgment, the very thing that makes human–AI partnership valuable, erodes from disuse.

THE GAP

Organizations need AI that is **capable enough** for real work, **small enough** to own and run locally, **honest enough** to show its sources and admit its limits, and **resilient enough** to work with no internet at all. That combination is what NanoSI is designed to deliver.

03

HARAT REVISITED

Adaptive, reflexive, agile

The 2022 conclusion defined three human systems. Read in 2026, they are also a specification for machines that work beside people.

HARAT's conclusion described an **adaptive** system that blends "one's mindset and skillset to achieve high probability target goals"; a **reflexive** system that requires "being aware of one's thoughts and feelings" while acting "with high emotional intelligence, morals, ethics and civic responsibility"; and an **agile** system of "quick, precise, decisive thinking actions in responding proactively to volatile changes, especially during uncertain times."

An AI meant to partner with people should embody the same three properties, in machine form:

HARAT SYSTEM	IN PEOPLE (2022)	IN NANOSI (2026 INTERPRETATION)
Adaptive	Blend mindset and skillset to reach high-probability goals.	Small models built from scratch and **fine-tuned for a defined job**, rather than one generalist for everything.
Reflexive	Awareness of one's own thinking; act ethically and responsibly.	The model **knows its sources and its limits**: it cites them, says "not supported" when they are silent, and leaves an audit trail.

Agile	Quick, precise, decisive action amid volatile change.	**Fast, local, offline** intelligence that keeps working through outages and in places the network never reaches.

The five co-generative principles

Each of the workshop's five principles maps onto a concrete NanoSI design decision.

#	PRINCIPLE (2022, CONDENSED)	WHAT IT BECAME IN NANOSI
1	The relationships among inputs are valued more than following steps and procedures.	**Structured knowledge:** facts are organized with addresses and relationships (document › section › page), not dumped in as loose text.
2	Optimizing a working solution for quick decisions is valued more than re-narrating the problem.	**Task-specific small models** that return usable answers fast, on the device.
3	Deliberate, dynamic responses to change are valued more than rigid procedures.	**Offline resilience:** the system keeps running when conditions change, in backup mode or permanent mode.
4	Continuous high-value, weighted information is valued more than random, unexamined inputs.	**Curated, source-linked knowledge** instead of internet-scale noise; answers come from what has been vetted.
5	Iterative fine-tuning of neural architectures is valued more than a single input.	**The ai021 method:** build from scratch, fine-tune for the job, test before it ships, iterate.



SECTIONS 04 – 07

PART II

The device came next.

How NanoSI turns the theory into architecture, and what is already built and running offline.

04

ARCHITECTURE

From theory to architecture: NanoSI

NanoSI is small enough to run offline and designed to show its work. It has three layers and one loop.

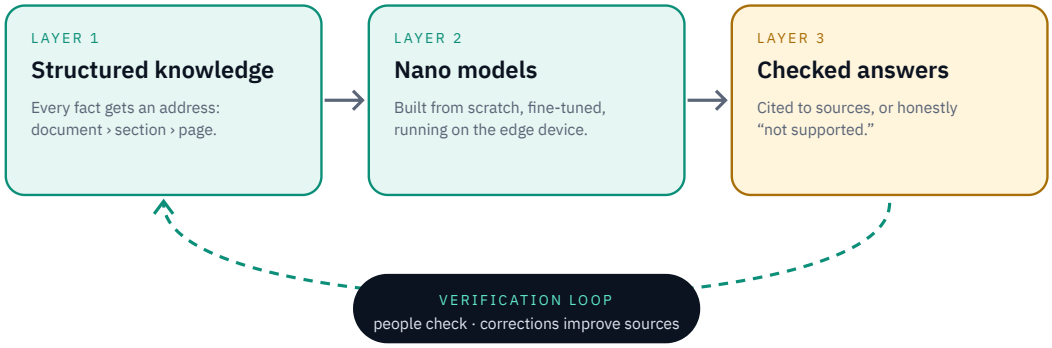


Figure 2. NanoSI: three layers and the verification loop, the concrete form of HARAT's "recursive iterative feedback loops."

Layer 1: Structured knowledge

Source material, such as manuals, procedures, policies, course content and specifications, is organized so that every fact has an address you can look up: the document, the section, the page. We picture this address space as a tesseract, where any point can be located precisely. Grounding a model in real source text at answer time is an established technique (Lewis et al., 2020). NanoSI makes it a first principle rather than an add-on.

Layer 2: Nano models

Small language models, built from scratch and fine-tuned for a defined task, run on the edge device itself. Fine-tuning mostly shapes *behavior*, how the model answers, while the structured knowledge supplies the *facts*. This separation keeps models small, keeps knowledge current without retraining, and

keeps the whole system able to run on modest hardware with no cloud call.

Layer 3: Checked answers

Every answer links back to its sources, and any answer can be walked back to the lines of the documents behind it. When the sources are silent, the right answer is “not supported,” not a guess. This is HARAT’s reflexive system in software: the model is aware of what it knows and what it does not.

CLOUD CHATBOT

“Can your AI guarantee every answer is correct?”

“Yes. Our model is highly accurate and you can rely on it for any question.” No source. Impossible to check. Online only.

STICKYSTUDIO · ON DEVICE · 1 SOURCE

Same question

“No. It can only verify what its sources contain, so every answer comes with the source you can check.” [1] ai021.ai › technology › known limits.

Figure 3. Illustration of the intended behavior, using a statement published on ai021.ai. Live behavior is shown in demonstrations.

05

INTELLIGENCE AT THE EDGE

Intelligence at the edge

HARAT foresaw intelligence becoming embedded and ambient. NanoSI puts it in a device you can hold, and keeps it working when the internet doesn’t.

More than a backup: a permanent solution

When the power fails, a generator keeps the lights on. The NanoSI device plays that role for AI and teamwork, and then goes further.

BACKUP MODE

Takes over when the internet goes down

Meetings, work and AI keep running through outages, then carry on when the connection returns.

PERMANENT MODE

The everyday system, offline for good

Runs all day, every day where there is no internet, where it cannot be trusted, or where it is not allowed.

Privacy, security and IP protection by architecture

Most privacy promises are policies. NanoSI's is physics: if the device is not connected to the internet, the data on it is not sent anywhere. Documents, conversations and know-how stay on hardware the customer holds. Designs, contracts, code and trade secrets are not uploaded to a third-party AI, so they cannot end up in someone else's training data. The device can be unplugged, locked away, or run on a closed, air-gapped network.

	TYPICAL CLOUD AI	NANOSI DEVICE
Where your data goes	The provider's servers	**Stays on the device**
Works with no internet	No	**Yes**
Could your data train someone else's model?	Depends on the provider's terms	**It isn't sent anywhere**
Who controls access	The provider	**You, physically**
Answers traceable to sources	Varies	**By design**

Why now

Three developments turned edge AI from a research idea into a buildable product:

- **Small open models became capable.** Openly documented small models, such as the SmolLM3 family (Hugging Face, 2025), show what compact models can now do.
- **Running models on machines you control is documented and practical,** with published tooling such as Ollama and vLLM (Kwon et al., 2023).
- **Smaller, well-trained models can beat larger ones on a defined task** (Hoffmann et al., 2022; Ouyang et al., 2022).

Together with demand from regulated industries for AI they can audit, these shifts make HARAT's "embedded, distributed, situated" intelligence practical for the first time.

06

PROOF POINTS

The first device and the Sticky suite

A theory earns trust when it ships. ai021's first edge device is built, and three products run on it today. DEMO READY

The first NanoSI device is a **portable USB unit** that runs ai021's own small language models **completely offline**, for conferencing and digital work. Connected to a local router, it is designed to serve **300–500 users at once**. That capacity is projected and will be confirmed in load testing before it is presented as a measured result.

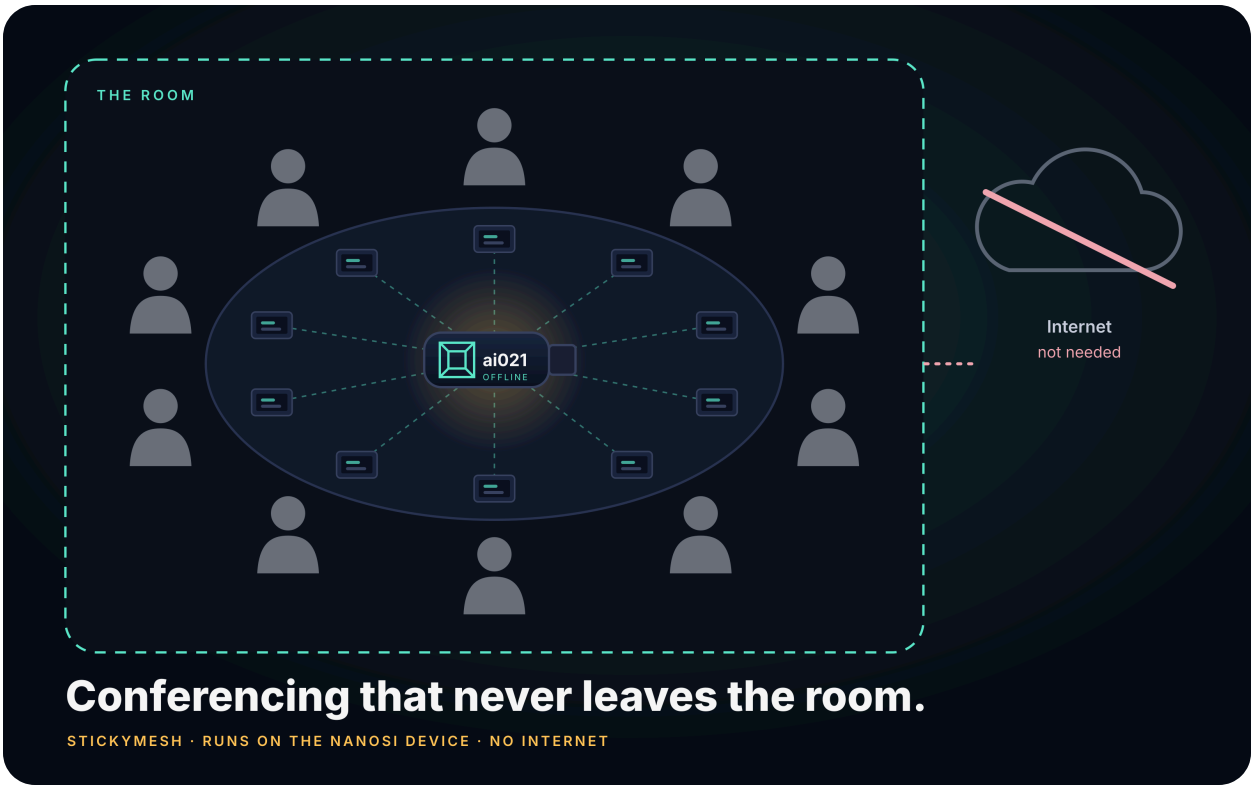


Figure 4. StickyMesh: conferencing that never leaves the room. Illustration, not a screenshot.

PRODUCTIVITY TOOLS

StickyWork

Everyday digital work, fully offline, on the same device as StickyMesh and StickyStudio.

COLLABORATION & CONFERENCING

StickyMesh

People meet and collaborate over a local router with no internet and no outside cloud service in the loop.

SLM INFERENCING

StickyStudio

Inference on small language models built and fine-tuned with ai021's method, on the device. NanoSI's source-linked answers in practice.



Figure 5. StickyWork, saving to the device while offline. Illustration.

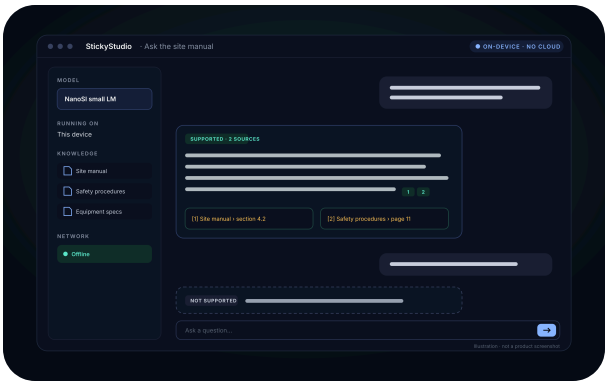


Figure 6. StickyStudio answering from local documents with sources. Illustration.

Early evidence: small and specific beats big and general

In a test published in ai021's lesson series (*"The boring baseline wins"*), we classified 77 kinds of banking questions two ways.

92.9%

Small embedding model plus logistic regression, at about 12 ms per message.

47.1%

Small chat model, prompted but never trained on the job, scored on a test slice.

The lesson is the one HARAT taught educators: a well-designed, focused learner beats a vague generalist on a defined task. It is a small, single-task result, not a general benchmark, and we present it that way.

What has been built

MILESTONE	STATUS
A method to build small language models from scratch	**Built**
Fine-tuning small models for a defined job, tested against it	**Built**
A working offline edge device (portable USB)	**Built · demo-ready**

StickyWork, StickyMesh and StickyStudio running on the device	**Demo-ready**
300–500 concurrent users over a local router	Projected · load testing planned

07

LEARNING, APPLIED TO MACHINES

The learning hierarchy, applied to machines

The 2022 workshop grounded HARAT in how people learn. ai021 applies the same science to how its models are built and how they work beside people.

From acquisition to adaption

The workshop drew on Haring and Eaton's learning hierarchy (1978): new skills move through acquisition, fluency, maintenance, generalisation and adaption. ai021 uses the same sequence as a lifecycle for its models.

STAGE	IN A LEARNER (HARING & EATON)	IN A NANOSI MODEL
Acquisition	The skill is learnt to accuracy.	Build the small model from scratch and give it a domain's structured knowledge.
Fluency	Performed accurately and fluently.	Fine-tune for the job; test for accuracy and speed on the device.
Maintenance	Retained and retrievable.	Held-out tests, updates when sources change, checks for drift.

Generalisation	Applied to different situations.	Deploy to new sites, teams and settings on the same device.
Adaption	Adapted to new conditions.	Carry the method to new edge hardware and, next, into an operating system.

The zone of proximal development, for teams

Vygotsky's zone of proximal development describes the space between what a learner can do alone and what they cannot do at all, the tasks they can master *with guidance*. HARAT's view of AI fits that zone: not a replacement that does the work, but a scaffold that lets a person reach further while their own judgment grows. That is why NanoSI shows its sources. A scaffold you can inspect teaches; a black box only answers.

Co-creation, not substitution

The workshop drew a sharp line about what human–AI co-creation is not. Among the items it ruled out was *"AI working alone in decision optimization."* The same rule runs through ai021's product principles: people keep authority over decisions, the system records what it relied on, and the human stays in the loop by design.

> *"The human mind and AI can achieve better results when collaborating with a shared purpose."*
>
> KHAN, HARAT, 2022



SECTIONS 08 – 11

PART III

The road ahead.

Where it matters, how trust is designed in, the roadmap to an AI operating system, and what it means for investors.

08

WHERE IT MATTERS

Where it matters

The places that most need trustworthy AI are often the places cloud AI serves worst.

SETTING	THE SITUATION	WITH THE NANOSI DEVICE
Oil rigs & offshore	Limited, expensive satellite links that drop in bad weather; crews still coordinate shifts and briefings.	Briefings over the platform's own network (StickyMesh), shift work offline (StickyWork), on-board AI with no data leaving the site (StickyStudio).
Schools	Slow or unreliable internet; some cannot send student data to outside AI services.	A classroom or school on one local network; lessons continue during outages; AI help that stays inside the school.
Business continuity	An outage stops calls, shared work and cloud AI; every hour costs money.	Switch to the device and keep meeting and working; client files and trade secrets stay on hardware you control.
Remote & field sites	Mines, farms, construction and research stations far from reliable coverage.	Plug into a site router and the team is connected; work is recorded on site; AI where the work happens.
Emergency response	After storms, floods or fires, the internet is often the first thing to go.	A portable local command network set up fast; coordination with no outside connection.
Secure facilities	Government, defense, legal and healthcare settings may forbid outside services.	Everything runs on the device on a closed network; answers traceable to their sources.

Table: example scenarios showing where offline AI fits. Not customer case studies.

Micro, meso, macro

The workshop's task-design "iceberg" looked at three levels, from partnerships to communities to nations. NanoSI scales the same way:

MICRO

A team in a room

One device, one router: a crew, a classroom, a project team working privately and offline.

MESO

A site or community

Schools, platforms, clinics and offices with their own local AI and collaboration layer.

MACRO

Resilient infrastructure

The long view: intelligence that a region or nation can run on hardware it controls, independent of distant data centers.

09

TRUST BY DESIGN

Trust by design

Trust is not a marketing claim. It is a set of habits: identify risks, control them, review the controls, and say plainly what the system cannot do.

A risk loop borrowed from the classroom

The 2022 workshop closed its task-design material with a simple risk-management cycle: **identify** risk, **assess** risk, **control** risk, **review** controls. ai021 applies it to every deployment.

STEP	APPLIED TO A NANOSI DEPLOYMENT
Identify	Which data will the device hold? Who will use it? What happens if a source is wrong or missing?

Assess	Which answers carry real consequences (safety, money, compliance) and need the strictest sourcing?
Control	Limit the model to vetted sources; keep the device offline or on a closed network; keep people in authority over decisions.
Review	Use the audit trail to see which facts and documents answers relied on; update sources; re-test.

Known limits

We publish these because investors and customers deserve to know them:

- Results depend on document quality, source coverage and domain.
- The system can only verify what its sources contain.
- There is no guaranteed performance level.

Values: the HARAT core, applied to a company

The workshop named five core values for teaching and learning excellence. They are how ai021 intends to operate.

VALUE	WHAT IT MEANS AT AI021
Courage	Building small, local and private when the industry is betting on big and centralized.
Focus	One job per model, done well, rather than one model for everything.
Commitment	Shipping working hardware and software, then proving claims before repeating them.
Respect	For users’ privacy, ownership and judgment. Their data stays theirs.
Openness	About sources, limits and what is still a plan.

10

ROADMAP

From one device to an AI operating system

The first device is a beginning, not a destination.

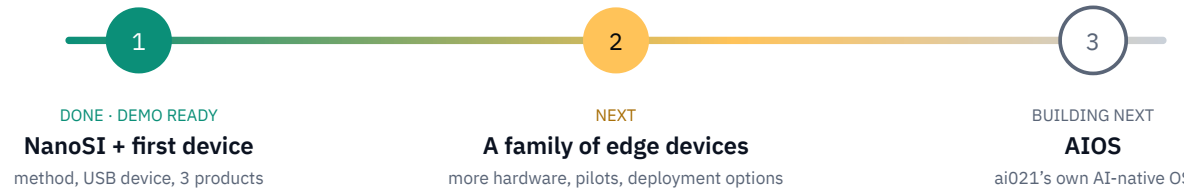


Figure 7. Roadmap. Stages 2 and 3 are in planning and development, not available products.

Stage 1: NanoSI and the first edge device (done, demo-ready)

A method to build and fine-tune small models from scratch, running offline on a portable USB device with three demo-ready products.

Stage 2: a family of edge devices (next)

The same method on more edge hardware, plus pilots in places that cannot count on the internet. Planned deployment options also include the customer's own cloud, on-premise servers and air-gapped networks.

Stage 3: AIOS, ai021's own AI-native operating system (building next)

An operating system for desktop and edge devices in which intelligence is a native part of computing rather than an app layered on top. It is designed around persistent memory, specialized AI agents and safety controls, with human authority and privacy at its core. AIOS is the fullest expression of HARAT:

adaptive agents, reflexive safeguards and agile, local intelligence, all answerable to the person using them.

UNDER NDA

The detailed technical approach, security architecture, capacity testing plan, go-to-market plan and use of funds are shared with qualified investors under a non-disclosure agreement, together with a live demonstration of the device.

11

WHAT THIS MEANS

What this means

For investors and venture partners

- **A category with structural tailwinds:** private, offline, auditable AI for settings that cloud AI serves poorly, from offshore operations and schools to regulated and secure organizations.
- **Built, not promised:** a working device and three demo-ready products, with claims labeled honestly where testing is still ahead.
- **A long-horizon platform:** from a single device to a family of edge hardware and an AI-native operating system.
- **A founder-led thesis with a public paper trail** that predates the generative-AI boom, from a researcher and educator whose career has been about how people learn and how to check that learning worked.

For organizations considering a pilot

The workshop's change model, ADKAR (Hiatt, 2006), is a practical guide to adoption: **awareness** of the need, **desire** to change, **knowledge** of how, **ability** to do it, and **reinforcement** so it sticks. Its task-design builder (initiation, planning, execution, control, closure) is how ai021 structures a pilot: choose

one narrow, repeated job where every answer must be checked, define the test first, deploy on the device, measure, and decide.

For educators and AI enthusiasts

HARAT predicted demand for “Human Expert Feature Engineers,” people who shape what machines learn. Today that role looks like teaching, and HARAT’s message is more relevant now than in 2022. Use AI to reach further, not to stop thinking. Ask where an answer came from. Teach machines well, because the skill that matters next is not typing clever prompts; it is explaining clearly, checking the work and correcting mistakes, which is what good teachers have always done.

CONCLUSION

Reason, then act

The 2022 workshop opened with W. H. Auden’s warning that refusing to reason and refusing to act are equally fatal. HARAT was the reasoning: people and machines can think better together, as long as human judgment stays at the center and machines stay honest about what they know. NanoSI is the action: small models, built from scratch, running on a device you hold, answering from sources you can check, and working whether or not the internet does.

The workshop closed with a hope, voiced by Max Tegmark, that amplifying human intelligence with AI could help civilization flourish, *“as long as we manage to keep the technology beneficial.”* ai021 exists to make that condition an engineering requirement rather than a wish.

ABOUT

About the author

Zafrullah Khan, Ed.D., is a futurist and AI researcher, and the founder and CEO of ai021.ai, Inc. He came to AI after thirty years in education. He taught in Durban, South Africa, and at Tiospa Zina Tribal School in South Dakota; was a school psychological examiner in South Dakota (1995–1998); led special education for Williamson County Schools; directed research for the Tennessee Department of Education; and prepared teachers at Belmont University and Middle Tennessee State University, where he was an associate professor from 2005 to 2022. He holds an Ed.D. from the University of South Dakota, a master's in special education from Minot State University, and a B.A., Higher Diploma in Education and B.Ed. from the University of Durban-Westville (now the University of KwaZulu-Natal). In 2022 he presented the peer-reviewed HARAT workshop at the 20th Annual Hawaii International Conference on Education. He is the author of nine books on building, using and trusting AI, including *Train Your Own LLM*, *Elastic Intelligence* and *Who Will Teach the Robots?*

About ai021.ai, Inc.

ai021.ai, Inc. (founded 2026) builds NanoSI: small language models that run offline on edge devices, with answers that can be traced back to their sources. Its first device, a portable USB unit, runs three demo-ready products: StickyWork, StickyMesh and StickyStudio. The company is founder-led; only the founder is shown, and there is no anonymous team. **Book a live demo or request the investor overview at www.ai021.ai.**

References and notes

Haring, N. G., & Eaton, M. D. (1978). Systematic instructional procedures: An instructional hierarchy. In N. G. Haring, T. C. Lovitt, M. D. Eaton, & C. L. Hansen (Eds.), *The fourth R: Research in the classroom*. Charles E. Merrill.

Hiatt, J. M. (2006). *ADKAR: A model for change in business, government and our community*. Prosci.

Hoffmann, J., et al. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems 35*. arXiv:2203.15556.

Hugging Face. (2025). SmolLM3 model card. huggingface.co.

Khan, Z. (2022). *Beyond critical thinking and creativity: Human Adaptive Reflexive Agile Thinking (HARAT) using different combinatoric neural network architectures for supporting human decision optimization when working with artificial intelligence (AI) in education* [Workshop]. 20th Annual Hawaii International Conference on Education, Waikoloa, HI. Edited edition (2026) available at www.ai021.ai/research.

Khan, Z. (2026). The boring baseline wins [Lesson]. ai021.ai, Learn series.

Kwon, W., et al. (2023). Efficient memory management for large language model serving with PagedAttention. *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*. (vLLM.)

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems 33*.

Ollama. Open-source tooling for running language models locally. ollama.com.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35*.

Tegmark, M. Quoted in the original 2021 HARAT proposal; the quotation is attributed to Max Tegmark, President of the Future of Life Institute.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Notes. Khan (2022) quotations are from the edited 2026 edition, which keeps the original wording. Auden's poem is paraphrased, not reproduced. ChatGPT's November 2022 release is cited as industry context.

ai021

WHITE PAPER NO. 1

FOR INVESTORS AND PARTNERS

See it run.
Offline, in the room.

The first NanoSI device and the Sticky suite (StickyWork, StickyMesh and StickyStudio) are ready for live investor demonstrations, with no internet connection required.

Private by architecture. Data, prompts and models stay on a device you physically control.

DEMO Book a demo · www.ai021.ai

READ The 2022 HARAT workshop materials · www.ai021.ai/research

© 2026 ai021.ai, Inc. All rights reserved. Product statements describe work in development. They are not promises or an offer of securities.

AI021.AI, INC.