WHITEPAPER

The Zero2AI Network

A decentralized global platform for Trusted AI.

Version 1.07 June 2020

Contents

Abstract 3

State of AI 4

Zero2AI 5

Architecture 6

Automated & Secure 6

Scalable 7

Immutable 7

Trusted 7

Decentralized 8

Resilient 9

Future 9

References 9

Abstract

Artificial Intelligence (AI) powered by Machine Learning (ML) may be one the greatest enablers of innovation the world has seen since the advent of the Internet. As of today, AI augments human capacity to automate cognitively routine tasks, enabling the broader human enterprise to focus on fundamental innovation. It is inevitable that in the near-future AI will assist in novel discoveries and drive core innovation across many industries. Already AI transcends human capacity in narrow expert-dominated fields like healthcare, manufacturing and financial markets. Considering the time saved from automating and accelerating tasks that involve human intervention, widely deploying AI can amplify our collective cognitive capacity as we attend to critical issues that plague civilization.

Zero2AI proposes a global AI network that makes deploying and scaling of AI solutions easy for the AI community. We believe that the best way to unleash AI innovation is to make AI accessible to the broadest possible audience. Zero2AI will lead the way as a trusted platform for AI consumption worldwide.

Combining decentralized compute and storage, the network aims to provide an AI-as-a-Service platform which will allow model publishers (software developers, machine learning engineers and data scientists) to effortlessly deploy, manage and scale their ML models to a global audience with minimum cost, guaranteed SLAs and immutable transparency.

The network will also provide a marketplace for Trusted AI. Using blockchain technology, Zero2AI aims to provide an immutable decentralized ledger that will underpin all operations on the network while providing an incentive-driven collaboration mechanism that will push trusted and socially responsible AI models to the top.

And finally, the network incorporates best practices of military-grade security for AI. Using a combination of encrypted execution and delivery, adversarial-defense testing and load/scale testing, Zero2AI will ensure that all models on the network are secure, resilient and scalable.

State of AI

The AI landscape appears to be fractured. On one end we have seen breakthrough innovation in AI algorithms, especially around Deep & Reinforcement Learning making AI models outperform humans in a wide swathe of tasks. This field is actively researched, funded and progress is happening at an accelerating pace. The bets on reaching the singularity are any one's guess, but at the current rate of investment and innovation AI will soon be nestled into every modern information workflow, from factories and hospitals to offices and homes. And this is for good reason, humans augmented with AI will only push the species forward and it is essential that we collectively understand and contribute to this phenomenon of creating post-evolutionary intelligences.

However, on the side of AI delivery the picture is far from rosy. For aspiring entrants, independent developers or cash-strapped startups who want to contribute to the AI vision, the inhibitions to AI innovation are manifold. They can be broken down across three dimensions:

Siloed AI platforms:

- Concentrated power in the hands of a few monopolies
- Potential lock-in with cloud platforms and tools
- Cloud, Edge and IoT suffer from a fragmented landscape.
- Intelligent Pooling of hardware is missing, limiting scale.

Lack of trust and accountability:

- No repeatable, long-lived AI models with transparency and attribution
- No transparent incentive structures built-in for validation and contribution
- No structure for AI model as global good

Barrier to entry – cost and ease

- Requires deep knowledge – not easy for non-experts
- Production deployment hard – scale, diversity, security, and maintenance
- No easy path to production and commercialization for the community
- Limited and specialized hardware leads to higher costs

At its core, AI models are software that once built need to be tested, packaged, delivered, executed, monitored, debugged, audited and updated - all of this while keeping security at the forefront of every workflow. What is needed is a trusted open platform for execution of AI models which will incorporate all these features.

Zero2AI

The Zero2AI network is a decentralized AI SaaS that makes testing and deploying AI models easy at global scale without lock-in and without sacrificing security, control or observability. Once models are ready to test at scale or deploy to production it is handed over to Zero2AI and the service takes over. An automated pipeline with auto-hardening and built-in benchmarking accelerates production delivery. The SaaS platform is engineered on a global network of decentralized compute and storage.

A key concept in Zero2AI is the notion of Trust as it pertains to an AI model. Trusted models are validated in an incentive-driven and distributed fashion by pre-selected nodes on the network. A model's Trust score is multi-dimensional, encompassing prediction performance on unseen real-world samples, prediction throughput and latency during massively parallel global requests, resilience to adversarial attacks and the model's potential impact on society in alignment with the principles of responsible AI. The provenance of model trust is encoded in an immutable distributed ledger called the ZeroChain.

Zero2AI CDN creates resources on-the-fly as demand oscillates, both within local geographies as well as remote locations as a reaction to geo-specific user requests. It provides autonomous, learned, anticipatory scaling for the Zero2AI network. The model publisher can specify budget guidelines, affinity, and additional custom constraints to optimize model delivery in the target geography. Zero2AI CDN also provides auto-healing and tenant specific requirements to meet stringent SLAs.

The Zero2AI network can securely deliver and execute across global compute farms. The network aims to leverage compute providers (including private clouds) worldwide both to enable decentralized AI delivery and lessen the dependency on cloud monopolies. Secure encrypted delivery allows models to run on untrusted hardware. ZeroChain ensures immutability, audit-trails, and accounting wherever a model runs. Zero2AI also replicates models and code securely on an underlying decentralized storage network. Even through a large global public pool is conceived, enterprises, institutions, and consortiums may carve out their own private global AI pools with access control and customized SLAs.

Zero2AI aims to reach the widest possible audience at the lowest possible cost. This requires leveraging under-utilized specialized hardware but also commodity hardware. Models deployed on the network are automatically optimized/compressed to support all hardware available from high-end NVIDIA hardware to AMD GPUs, ARM at edge locations, IOT hardware even smartphones for optimal end-user experience securely on low power edge devices.

Models deployed on Zero2AI automatically gets load tested to establish robustness and benchmarked for performance. The intent is to make the deployment experience painless for end-users without sacrificing the rigor of global deployment. The delivery pipeline is designed from scratch to support AI model specific hardening steps like domain specific adversarial testing.

Architecture

The Zero2AI Architecture is split along six essential capabilities.

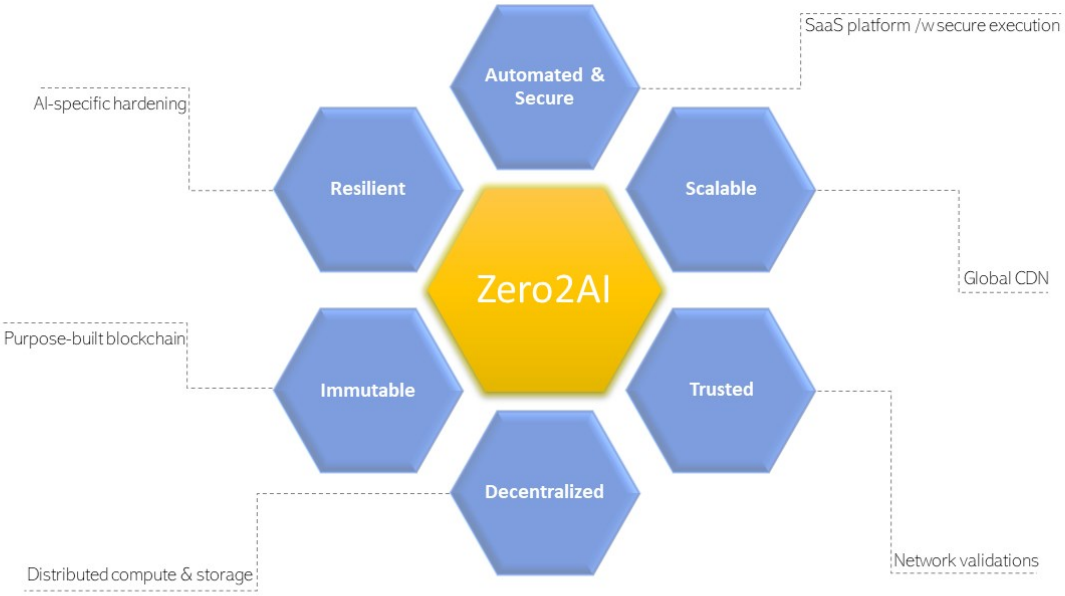

Diagram 1: Zero2AI Architecture

Automated & Secure

AI SaaS: A secure multi-tenant access-controlled SaaS that serves as a one-stop user experience for model developers, testers, validators and public model benchmarks. A signup and pay-per-use experience which allows use-case specific customizations, providing a near management-less AI platform.

Secure model execution: This facilitates distribution of model artifacts and running inferencing across untrusted hardware with non-repudiation and authentication. Zero2AI automatically leverages encrypted memory where available to run containers and virtual machines with unprecedented privacy and anti-snooping. Secure enclaves (like Intel SGX) will be employed wherever available.

Auto-detection of model domains: Model domains refer to data type that is relevant for the model, for example: text, numeric, video, audio etc. Each of these require bespoke setup to properly consume data and generate the right output. Zero2AI automatically detects model domains and sets up model preprocessing and hardening pipelines.

Model deployment automation: Updating models, A/B testing, remote wiping models either user directed or autonomously based on policy encoded as smart contracts. Publishers will be able to use the smart-contract abstraction to encode policy around model deployment events.

Scalable

Model DNS: Expanding on the well understood concept of Domain Name Service on the Internet which allows users to find web resources through a distributed translation mechanism, Zero2AI will create a mechanism for models to have a public naming scheme for models to be found and translated to an API end-point at runtime. Using a decentralized approach, a mechanism will be available for establishing the right endpoint to contact for model execution. Depending on a user's geo-location a lightweight intelligent mediation service will translate and in real-time redirect a user to the right potentially local endpoint. This will involve secure authentication and encrypted communication.

Immutable

ZeroChain: One of the core components of the Zero2AI network is a specialized distributed ledger which underpins all network operations, records trust and data provenance, and provides the backbone for trusted, transparent, and socially responsible AI models. This ledger is based on blockchain technology and can potentially be layered on and extend public blockchains like Ethereum.

Model versioning, ongoing benchmarking, contributors and compute providers are immutably recorded on ZeroChain and surfaced via dashboards that are easy to understand from the AI practitioner, AI consumer and administrator point-of view.

Trusted

Built on ZeroChain, the network aims to provide a framework where certain network participants, called validators, can collaboratively evaluate the prediction performance of any Machine Learning model. This decentralized validation increases baseline model performance – a model benchmark which reflects the cumulative effects of network validations over time - leading to greater confidence amongst consumers. The validation is driven by an incentive mechanism built into the framework which reimburses model validators with network tokens proportionate to their contribution in improving the model.

Model Validators and Consensus: Participants in the network who have been authorized by the model publisher to validate the efficacy of the model. In a private organization the validators are model-testers, QA or even non-human agents running validation code based on certain rules. In the public incarnation of the network, the validators could be elected based on decentralized voting scheme.

Model validators download a newly published model, run validation code and generate performance metrics e.g. precision, recall & F1 score. A network client maps these metrics to a scale and based on certain thresholds sends a positive or negative endorsement (plus additional information that captures validation specifics) signed by the validators digital signature.

The act of model validation and the quality and quantity of validation itself is used as a consensus mechanism. The models themselves are non-fungible tokens once minted.

Model benchmark: The model benchmark produces a score which quantifies the quality of the model i.e. its prediction performance when tested against diverse, non-biased and unseen samples. The datasets used for the validation are generated by the model validators and must meet certain criteria before they can be used i.e. the datasets must:

1. Be free from known biases.
2. Be purged of any private information/anonymized.

3. Not align to the distribution of the test dataset in the model artifact (i.e. no overlap with the data that was used by the model publisher).
4. Include adversarial-attack samples.
5. Stress-test the model for compute time performance.

The score is computed on the network once the validations and validation metadata are broadcasted by the validators. A cumulative algorithm is used to calculate the score which will change to adapt to an increasing number of validators and/or consumers of the models. The model benchmark is the only assurance that a model consumer has that the model is (a) performant (b) unbiased (c) secure and (d) will scale to global requests.

Incentive Mechanism: In a private organization, the incentive mechanism increases reputation for parties that validate a performant model. Validation based consensus allows these reputable validators to rise to the top of the hierarchy which may map to a non-network compensation provided by the organization. On the public network, the incentive mechanism encourages a game-theoretic approach to competing for both reputation and network tokens which can be exchanged for fungible cryptocurrency in the future. A malicious user will be penalized for broadcasting positive endorsements for a weak model, either through a proof-of-stake system (where validators must initially provide a stake) or through a potentially irreversible reputation hit.

Model Auditability: ZeroChain registers all model operations – publish, update, predict – in an immutable distributed ledger. It also provides an interface to do regulatory audit for any given model or across the network. Who used which model and when? What were the results? Were models used in combination to affect a certain workflow? Who validated the models? What were the benchmarks when the results were achieved? These and many more questions can be answered via invoking the ZeroChain audit interface.

Decentralized

Distributed Compute: For compute the network enables onboarding of multiple compute providers (public and private) and edge destinations that are optimally used for model deployment based on configured metrics.

Decentralized Model Store: The distributed model store (potentially leveraging de-centralized object storage e.g. private IPFS cluster) is the underpinning for storing encrypted models, test data collected from user interactions and log & usage data. Durability, longevity and availability are characteristics of the network. De-centralization also insulates against dependence on specific providers. Strong encryption and authentication assure integrity of data and access only by authorized individuals regardless of whether objects are accessible/available on public networks.

The Zero2AI network gets out of the way of model execution once a model has been immutably published. For performance, scale and privacy reasons, the end-user connects and executes the model using a peer-to-peer architecture, much like BitTorrent.

For this model-compute-user handshake to occur, several resources need to be pre-orchestrated. First the Zero2AI DNS service translates and redirects to the appropriate local, public or private endpoint based on real-time information. The Distributed Model store needs to have enough replicas setup ahead of time. The Distributed Compute framework should have enough resources available to run the model inference code.

Zero2AI continually learns from these subsystems to predict network load and usage patterns globally to cache and pre-provision resources appropriately. Once an end-user receives an endpoint (through a secure handshake) the user directly interacts with a nearby compute provider (this could be the user’s end-device or smartphone) and starts getting prediction results.

Resilient

AI delivery with hardening: Modern security implies robust testing, immutability, and auditability. Modern security also demands hardening. All models published on Zero2AI automatically gets load tested to establish robustness and benchmarked for performance. The intent is to make the deployment experience painless for end-users without sacrificing the rigor of global deployment. The delivery pipeline is not a standard software engineering pipeline but designed from scratch to support AI model specific hardening steps like domain specific adversarial testing. Domain specific synthetic data is utilized which gets infused with real prediction data over time, automatically perturbed to test robustness of model prediction and receive a grade as to the ability of the model to withstand adversarial attacks. Zero2AI encourages the community to publish models, to illustrate concepts, collect data and improve models with the feedback. Early stage model or mature models both are supported with robust feedback and training data augmentation mechanisms as soon as models are published and start generating predictions based on world-wide usage.

Future

The last 10 years has enabled us to make a quantum leap in creating AI systems, but there is a long way to go, both in creating Artificial General Intelligence and in incarnating the narrow, domain-specific AI that has emerged out of the current wave[1]. By all accounts we are still in that wave, going by the unprecedented volume of cutting-edge research that is pumped out of academia, industry and independent researches alike. What is needed is a way to consume this wellspring of creativity and apply it to scenarios in our life.

Cloud Computing brought the concept of utility to compute power, delivering on-tap compute to any device connected to the Internet. Zero2AI intends to bring the concept of AI as a utility, a truly decentralized AI platform that can reach any device on the Internet. This will allow anyone to tap into domain-specific intelligence, served by the platform in the form of rudimentary AI models and eventually by autonomous intelligent agents which will provide a broader and deeper context into solving the problem at hand.

The universe can be thought of as an infinitely large information processing system, and in our humility we need to understand that only by creating, disseminating and embracing non-biological intelligence can we begin to make sense of our human condition. Zero2AI envisions a world awash with such intelligence and is determined to broadcast it to the furthest reaches of the planet.

References

[1] DARPA The Three Waves of AI: https://www.darpa.mil/attachments/AIFull.pdf