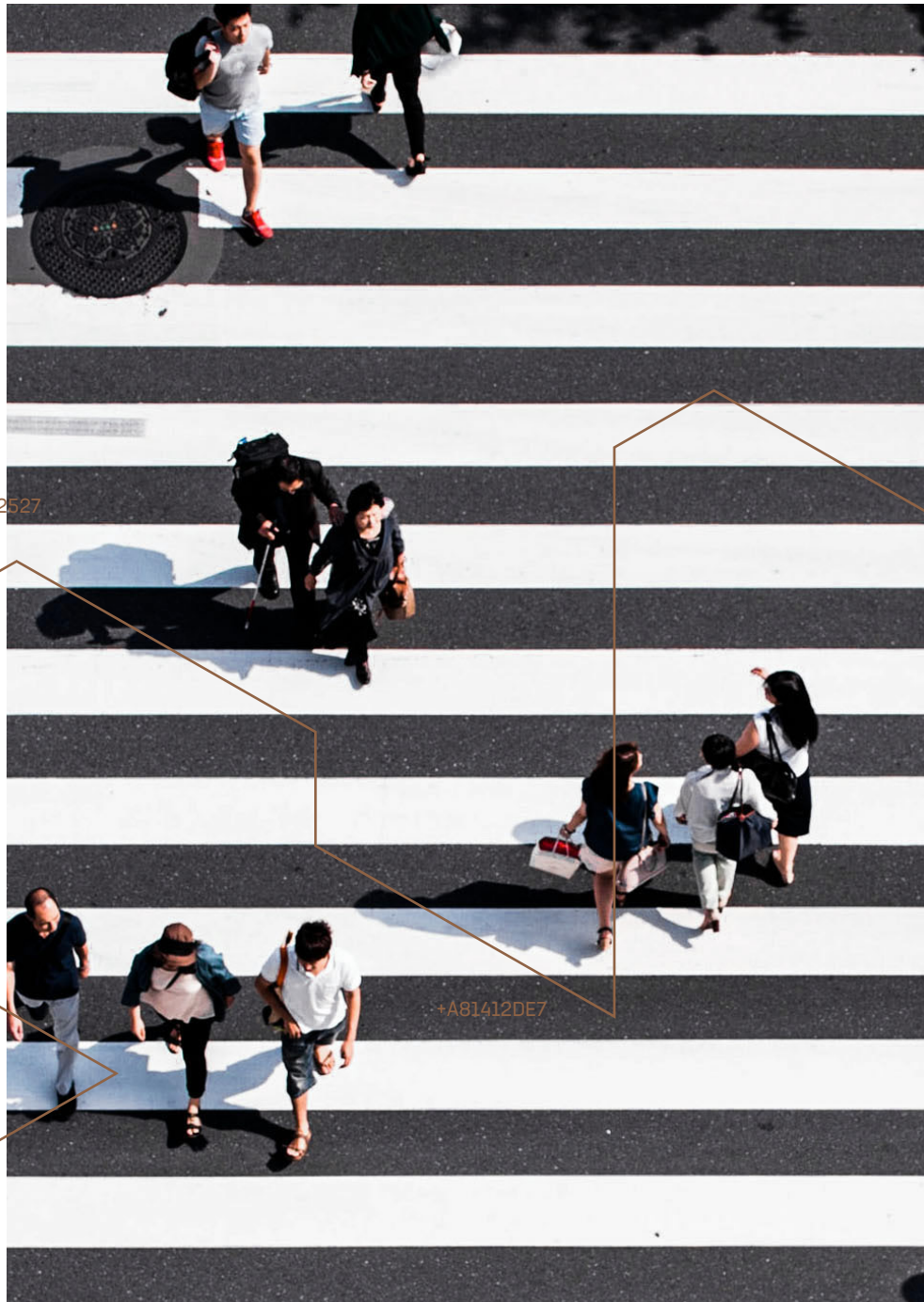
guardtime

guardtime.com

Data integrity critical in securing autonomous AI

by Guardtime Strategic Advisor
Tony Kenyon



AI, more specifically machine learning (ML), is a hot topic for every forward-thinking enterprise right now. The insights gained from analysing enterprise data are fast becoming a critical differentiator in enabling organisations to identify new opportunities, cut costs, and stay ahead of the competition. AI-driven systems are already demonstrating significant benefits to organisations willing to invest in them.

AI and blockchain are both cutting edge technologies, with the potential for significant value when combined, and today we see huge interest in the adoption of both, with spending expected to grow substantially over the next 2-3 years. IDC1 projects that annual global spending on AI is projected to reach $98 billion by 2023, and spending on blockchain solutions will reach US$16 billion by 2023. By combining blockchain and AI we have the potential to open up new and disruptive models for trusted data sharing data, enabling monetisation of new and untapped legacy data.

We have to consider that the ML systems we see today are becoming increasingly complex, with data moving offsite for cloud analytics, and new paradigms such as federated learning, with mobile, IoT and cyber physical systems, exposing much more of the ML infrastructure out to the edge.

We expect to see much greater autonomy, as processes become increasingly digitised and automated (so-called 'hyper-automation'), to the point where learning systems can directly act upon insights without human intervention.

Another key consideration for AI is the 'mission criticality' of emerging use cases; particularly where increased automation is expected across sectors such as industrial, energy, telecoms, intelligent transport and healthcare. We simply cannot afford to have intelligent systems take actions that have significant financial or operational impact, with the potential for life-threatening consequences, if those decisions are based on corrupt or poisoned data.

In today's increasingly regulated environment, data breaches happen almost daily, and we know that 'dwell time' for hackers can average several months, with plenty of time to cover their tracks. If we consider that many organisations are migrating towards entirely programmable infrastructures - often with AI in the decision loop - there is scope for major disruption, data tampering and data loss.

All these considerations mean that as ML continues to transform industry, organisations must take steps to ensure the integrity of their entire ML ecosystem, wherever components may be deployed. We need to address several key areas, including: safe data collection and storage, enabling data sharing and monetisation, assuring data provenance and consent with data handling, mitigating adversarial attacks, enabling explainable AI and immutable audit, and reducing subtle effects such as human bias and concept drift.

Data handling with immutable proofs

Successful application of ML's huge potential demands that systems collect, store and manage large volumes of data both reliably and securely, and today we see a clear trend where ML systems are integrated as a strategic function. These datasets and ML models often contain sensitive personal, financial and operational information, and the predictive outputs from these models can have significant external impacts. For these reasons ML systems are likely to be be attractive targets for cybercriminals.

It follows therefore that the security and integrity of the datasets and feeds used to train and inform ML systems, as well as the state of any parameter configurations, are critical threat surfaces, and naturally we will need to implement multiple layers of security, with firewalls, anti-virus, encryption, intrusion detection, and continuous monitoring.

However, what these traditional security controls often don't tell us is the state of the underlying assets, since controls are often designed to deal with external threats, and may rely on encryption as the primary means of ensuring integrity. As we know, encryption is an excellent way to ensure confidentially, however it does not provide provable integrity, nor does it scale particularly well.

In the case of AI it may not always be feasible to encrypt data using conventional techniques, and it may not be practical to run AI on encrypted data. Depending on performance or privacy needs, organisations may need to look at technologies like homomorphic encryption and differential privacy for example.

The key point here is that despite having mature and sophisticated security tools at our disposal, none of these address data integrity in a provable and scalable manner. This is where blockchain can really help. By leveraging the immutability properties of KSI, as well as the carefully design service infrastructure, we can register the state and provenance of all ML assets, in situ or in motion, and validate that state at any point in future, through the application of simple but rigorous cryptographic proofs.

Dealing with adversarial attacks

In addition to securing the ML data at rest, organisations need to be aware of more subtle issues inherent in ML learning process that can be exploited by adversaries: such as poisoning or disrupting existing telemetry feeds, and supplying false data. Perhaps the most prominent example is an 'adversarial attack', designed to mislead an ML system by supplying data that violates the assumption made on training data and skews the behavior of the ML model into producing incorrect insights, or taking bad decisions. Such attacks can be highly sophisticated.

There are a number of techniques to counter this behavior: two of the most effective being 'adversarial training' and 'defensive distillation'. Adversarial training injects such examples of adversarial attacks into training data to increase the robustness of the production model against such attacks. Defensive distillation ensures that an ML model is less susceptible to exploitation, by training one model to predict the output probabilities of a second model that was trained on baseline data. This makes it hard for adversaries to easily identity attack vectors to exploit. It's worth pointing out that neither of these techniques are foolproof, and in practice other techniques may be employed, such as ensembles. We could also consider techniques such as differential privacy on datasets to minimize the potential for sensitive data leakage.

Fundamentally we can mitigate some of these problems if we have visibility on the state of all underlying assets and telemetry feeds, full provenance on datasets, and tight access controls. This means instrumenting and securing datasets, the configuration state of edge devices, cloud infrastructure, providing end-to-end provenance across data feeds, and providing immutable audit trails. Given the scale and complexity of some of these infrastructures this is no easy task, however we can achieve this by leveraging KSI, with its unique ability to register and validate huge volumes of digital asset states in parallel.

Reducing the impact of bias

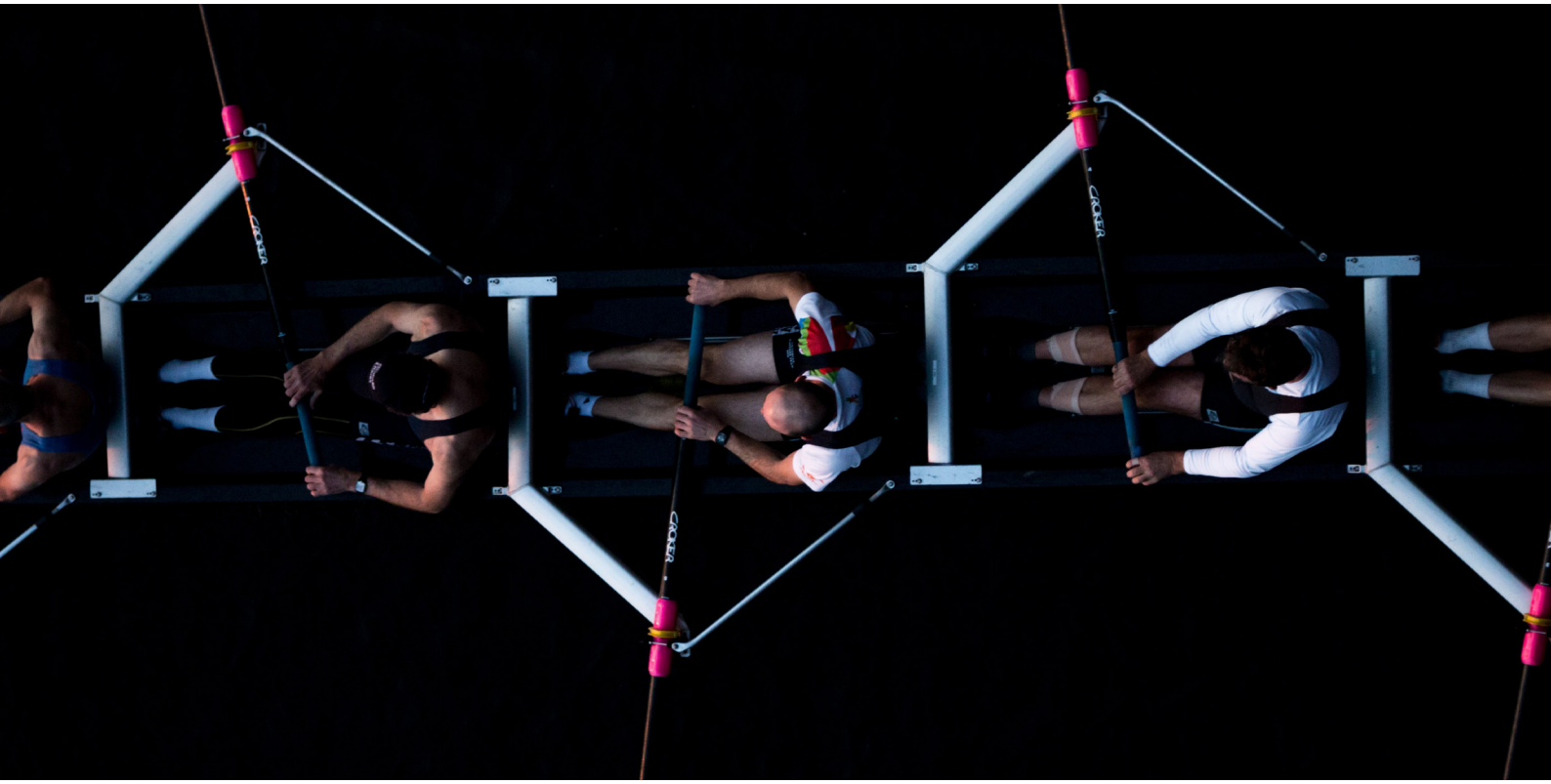
We might reasonable assume that predictions and decisions generated by an ML system are unbiased. Unfortunately this is not always true. At this point we should be clear what we mean by bias – what we mean here is that the ML is 'performing badly'. Essentially what we are talking about is the phenomena of observing results that are systematically prejudiced, due to erroneous assumptions.

Such bias can be engineered into ML systems if, for example, skewed, tainted or incomplete data is fed into business applications that the system operates on in production. By manipulating these inputs attackers may be able to influence the outputs of an ML system for their own benefit.

This can be partly mitigated with a sufficiently well trained system that is more immune to noise or can filter out anomalies. It may be possible to run different models as an 'ensemble', relying on a voting system to dampen erroneous outputs. In practice it may not produce a model that performs robustly when fed bad data, and we often refer to this as ML 'brittleness'.

A major obstacle for enterprises to combat bias is the lack of cohesion in data sources, with multiple legacy data silos that often emerging as the fallout from rapid cloud adoption. This can leave some aspects of the infrastructure with weaker access controls and protections. Automation and integration technologies that create more consistent end to end workflows can improve this situation, however we may still be left with heterogeneous datasets with no clear idea of state or the provenance of any subsequent data transformations.

Blockchain can help here by first securing the immutability of each data silo. We can go further by instrumenting data migrations, aggregations, and transformations that accompany these workflow processes, so that end-to-end integrity is registered and periodical verified against the blockchain - with clear provenance on origin data and subsequent transformation steps.



Securing permissions and consent

An increasing concern for ML models is the handling (or mishandling) of sensitive data, and whether adequate permission was sought to use that data for processing purposes. This is particularly important for long running projects in areas such a healthcare, where the benefits of using ML enable major advances in areas such a diagnostics, clinical trials, and retrospective document classification/annotation. Key issues here include the careless handling of PII information, whether consent was informed and granted for the purpose of processing, and whether appropriate de-identification techniques have been employed.

There have been a number of notable cases where ML has been used, often very successfully, but without informed consent from patients involved.

Given the direction of travel in regulation, particularly in the protections required on PII data with GDPR, it is imperative that datasets are unambiguously identified and tagged with such metadata, ideally with the identities of the associated authors of the dataset. Blockchain can help here is by providing an immutable registration and verification substrate for all of this metadata, as well as linking the specific identity of the dataset and associated individuals.

Dealing with concept drift

ML models are only as good as the data that feeds them, but they also rely on the set of assumptions made against that baseline data, and these assumptions can become brittle over time.

If operational data changes - perhaps legitimately, in response to environmental changes, or through malicious tampering – this can lead to a phenomenon called concept drift. Concept drift is essentially a discontinuity between a learning model and the current state of the environment. Concept drift can be engineered by feeding misleading data into business applications or poisoning sensor feeds, and many ML systems have limited ability to know whether incoming data is within acceptable bounds or suspicious. In some cases if the data is manipulated subtly over time it may be impossible to tell.

In effect, this means that a vulnerable ML model running in production could be operating on very different assumptions to the true underlying data, leading to potential misleading or suppressed insights, and bad predictions. Any subsequent actions taken on such insights may result in security exposure, service loss or degradation, data loss or corruption, even potentially life-threating actions, and these risks are amplified where systems are entirely autonomous.

Concept drift can be partly mitigated by periodic updating and retraining or learning models in production, or by using techniques such as adversarial training. In practice we need to be more proactive, mandating that the state of critical infrastructure be locked down, securing vulnerable end to end telemetry feeds, as well as instrumenting audit logs, based on periodic checking of any deviation in state against the blockchain.

The role of explainable AI

One of the key challenges in AI is the issue of transparency, and not all ML models are easily explainable (hence the association with the term 'black box'). When dealing with sensitive personal data in particular this can be a particular challenge in supporting regulations such as GDPR, where users may request further information on the nature of data processing.

Some ML models (such as neural nets) are effectively not much more than highly trained graphs of weights and vectors, based on large quantities of data, and as such not easily explainable. Such models may deviate over time and it may not be clear why. In many cases there may be no accepted techniques to properly understand model behavior, other than by observing it, and this remains an active area of research.

If the original training data is later decoupled or lost, then effectively we have a working ML system that we may not be able to explain the exact functioning of, or even how it got there, it just works well. Going forward this may be unsatisfactory when mapped against compliance needs.

An emerging area of interest on this problem is the field of 'explainable AI' (XAI). Explainable AI offers reasoning as to why ML system arrived at various outputs and predictions, and attempts to provide explanations of how it got there. With advances in explainable AI we may be able to mitigate some of the threats around concept drift, adversarial attacks, and intentionally planted bias, however this is still a relatively new field. Unsurprisingly XAI works best with AI models that are inherently explainable, and as mentioned earlier, some AI techniques remain inscrutable.

On a more practical level, blockchain can help make AI more coherent and understandable by immutable registering all data, variables and processes involved in a decision process, anchored in time. We can then trace back and determine why particular decisions were made by the model, with the assurance that those states have not be manipulated. We can also use many of the techniques mentioned earlier, using blockchain proofs against the state of infrastructure, datasets and models, as well as associated telemetry. Without unambiguous proof of the state of these underlying systems the whole ML ecosystem cannot be declared secure.

Emerging challenges with federated learning

Federated AI is a relatively new and interesting development, pushing AI compute out to the edge and offering the possibility of maintaining local data with that remote ML compute resource. This has potentially key advantages in maintaining data privacy. In practice federated learning is challenging to deploy, with multiple tradeoffs in scale, speed, and privacy, and to date has been used only with a limited set of use cases.

From a security, integrity and privacy perspective, federated learning introduces a great deal more complexity, and pushes both training data and learning models out to edge networks and mobile devices, some of which may have far less security than a more centralised cloud approach. On top of that the protocols required to maintain such a decentralised and potentially unreliable infrastructure can be quite complex.

The blockchain techniques described earlier can be used here in a similar way that we might approach an IoT or sensor network. In some cases it may be desirable to deploy blockchain asset registration at the edge devices themselves, to guarantee the state of remote datasets and learning models, as well as end-to-end provenance of any data transfer, to prevent adversarial attacks. This really depends on the specific deployment features.

With the federated infrastructure, centralised ML models and associated data and hyperparameters can also be included in a provenance chain of blockchain asset registrations, so that we have a complete change history for improved auditability, and to enable trusted rollback. KSI can be employed to register and validate vector updates so that the core model has blockchain-backed proofs against the updates received. Blockchain may also play a future role in promoting incentives for remote users to allow updates to be fed back.

Data sharing and monetisation of AI

By combining the AI and blockchain technologies we open up the possibility of sharing and monetisation of data. Monetising such data is a major revenue source for corporations such as Facebook and Google, and many organisations today have vast silos of information, sometime collated over many years, that could be extremely valuable to data science and domain experts across fields as diverse as healthcare, retail, and insurance.

By using blockchain-backed proofs, together with privacy and identity management techniques we can see disruptive new revenue models emerging out of the availability of such data. Blockchain is also likely to play an increasing important role in value distribution, through features such as smart contracts.

We might even consider using blockchain to encourage a broader distribution of datasets and algorithms, helping promote advancement across the wider AI community, through the creation of 'decentralized AI'.

Why are blockchain and AI such a natural fit

The applications of machine learning today are truly revolutionary, but implementations must be designed with great care – especially as we move towards greater automation and autonomy in decision systems. Blockchain offers important complementary features that add significant trust to the AI ecosystem.

As AI continues to permeate across business functions and critical systems, we need tools to assure integrity at scale. Although the combination of blockchain and AI is still a largely untapped field (despite this being an active area of research) by putting these two powerful technologies together we can start to realise the potential for using and managing data in novel and disruptive ways.

By using Guardtime KSI, solution providers can access simple, well-abstracted APIs, that enable AI ecosystems to interact with mathematically rigorous proofs about the state of data, model, and telemetry - all possible at massive scale.

References

1. IDC, "Worldwide Artificial Intelligence Spending Guide", Sep 2019.
2. IDC, "Worldwide Blockchain Spending Guide", Aug 2019.