AI TRUST INDEX

The New Era of Anti-Agent Security
Новая эра анти-агентной безопасности

**Traditional antivirus, firewalls and VPNs were built for malware.
AI agents use legitimate tools, APIs and web ports.
The biggest leaks now come from inside — through invited, over-privileged autonomous systems.**

An independent, measurable standard for trust in AI systems, automation agents, integrations and the organizations that deploy them.
First implemented and validated on BizDNAi — the platform that lives by its own rules.

ENGLISH VERSION
Starts immediately after this cover — scroll down or use your PDF reader's Bookmarks / Outline panel (usually left sidebar) for one-click jump to English section

РУССКАЯ ВЕРСИЯ
Начинается после английской версии (примерно страница 6) — прокрутите вниз или используйте панель Bookmarks / Outline в PDF-ридере (левая боковая панель) для быстрого перехода к русской версии

Developed by Rashid Kabzhanov • bizdnai.com • kabzhanov@gmail.com
Open-source auditor: github.com/Kabzhanov/ati-audit

ENGLISH VERSION

What is AI Trust Index?

AI Trust Index is an independent, vendor-neutral standard and scoring system designed specifically for the age of autonomous AI agents and complex automation. It measures how well an organization or product understands, manages and communicates risks introduced by AI systems that have real tool-calling power — access to files, databases, APIs, browsers and messaging platforms.

Unlike traditional security certifications that focus on perimeter controls or known malware signatures, AI Trust Index evaluates the new attack surfaces that appear when intelligent agents are given legitimate access to business systems. These surfaces include prompt injection leading to unintended actions, over-privileged tool permissions, RAG memory leakage across tenants, multi-agent privilege chains, and covert data exfiltration through standard web protocols that look completely normal to firewalls and EDR solutions.

Core Components of the Score

- Risk Accounting — Does the organization systematically identify and document risks specific to AI agents and automation (not just generic IT risks)?
- AI Usage Policy — Are there clear, published rules governing when, how and with what permissions AI agents may operate on company data and systems?
- Legal & Regulatory Compliance — Mapping to ISO/IEC 42001 (AI Management Systems), Kazakhstan AI Law (disclosure and consent requirements), and Personal Data Law 2026. Includes G13 sub-score for personal data protection readiness.
- Process Maturity — Are there repeatable processes for agent deployment, permission review, logging of every decision and tool call, and incident response specific to agentic behavior?
- Team & Governance — Does the organization have people with explicit responsibility and competence to govern autonomous systems, or is AI treated as 'just another tool'?
- User & Stakeholder Satisfaction — Measured through transparent communication of AI involvement and actual user trust in the systems they interact with daily.

Privacy by Design: All assessments are performed inside the client's own environment. Only the final score, G13 sub-score and high-level justification are published. Raw data, source code, internal documents and personal information never leave the premises.

Why Traditional Defenses Are No Longer Sufficient

Firewalls, VPNs, antivirus and even modern EDR solutions were designed to detect and block malicious code trying to enter or operate from outside. AI agents do not arrive as malware. They are explicitly invited through chat interfaces, API integrations, RPA tools and custom automation scripts. Once inside, they use completely legitimate protocols (HTTPS, WebSocket, gRPC, OData, REST) and appear as normal user or service accounts.

A well-intentioned but poorly isolated AI Sales agent connected to WhatsApp and Bitrix24 can be manipulated via prompt injection to dump entire customer databases. A 1C integration bot with broad OData permissions can be instructed to export warehouse data to an attacker-controlled endpoint — all through 'normal' looking API calls. Traditional tools see nothing unusual because there is no exploit, no dropped file, no C2 beacon in the classic sense. The agent simply does what it was given permission to do.

This is the core shift: we have moved from 'detect bad code' to 'govern autonomous agents that were given legitimate power'. AI Trust Index is the first independent mechanism to measure and communicate how well this governance is implemented.

For Users — What You Should Know and Demand

If you use AI assistants, chat-based automation, or any tool that connects an intelligent agent to your work data (email, CRM, 1C, documents, messengers), you are already inside the new threat model.

The risk is not that 'someone hacks your computer'. The risk is that the AI agent you or your company invited has more access and less oversight than you realize. A single over-privileged agent can read, summarize and transmit sensitive information through channels that look like normal operation.

Practical questions every user should ask vendors and internal IT:

- Does this AI tool or agent have an independent AI Trust Index score? Can I see the detailed justification?
- What exactly can this agent read, modify or send? Is there a clear, auditable list of its tool permissions?
- Where does my data go when the agent 'processes' or 'summarizes' it? Is it logged and can I audit every decision?
- If the agent is compromised or manipulated (for example through clever prompts), what is the blast radius?
- Does the vendor publish how they isolate agents between different customers and prevent cross-tenant data leakage?

You have the right to demand transparency. AI Trust Index exists precisely so that users can make informed decisions instead of blindly trusting 'we use advanced AI security'.

For Companies — Business, Regulatory and Reputational Risk

For any organization that deploys AI agents internally or offers AI-powered automation to clients, the question is no longer 'do we have antivirus and firewall?'. The question is 'how do we prove that our autonomous systems are trustworthy and contained?'.

Kazakhstan's new AI Law and Personal Data Law (2026) introduce explicit requirements around disclosure of AI interactions and consent. Non-compliance is not theoretical — it creates real legal exposure. Beyond regulation, a single high-profile data incident involving an AI agent can destroy years of trust with customers and partners.

Key business implications:

- Vendor Risk Management — Before connecting any third-party AI agent, automation bot or integration (especially those touching 1C, Bitrix24, amoCRM, WhatsApp or warehouses), require the vendor to publish their current AI Trust Index score and report.
- Internal AI Initiatives — Custom agents built by your own teams or contractors must be assessed. 'It works in testing' is not sufficient when the agent has production data access.
- Investor & Board Reporting — AI Trust Index provides a single, comparable, defensible metric that can be included in security posture reports, due diligence packages and ESG disclosures.

• Competitive Differentiation — Companies that can demonstrate high Trust Index scores for their AI systems gain an advantage when competing for contracts with security-conscious clients and in regulated industries.

• Incident Preparedness — Organizations with mature agent governance (logged decisions, clear permission boundaries, rapid containment procedures) recover faster and face lower regulatory consequences when something goes wrong.

For Developers & Architects — The New Threat Model

If you build, integrate or operate AI agents, multi-agent systems, RAG pipelines, automation bots or any LLM-orchestrated workflow that touches real business data or systems, the security paradigm has fundamentally changed.

The classic application security model (input validation, output encoding, least-privilege database accounts) is necessary but no longer sufficient. Agents introduce new classes of risk because they are goal-oriented and tool-augmented. They can be manipulated not only through code vulnerabilities but through natural language instructions that change their behavior at runtime.

Critical areas the AI Trust Index evaluates for developer teams:

• Excessive Agency — Does the agent have more tools and permissions than strictly required for its stated purpose? Can it read entire file systems, execute arbitrary commands, or access production databases when it only needs to answer questions about one module?

• Prompt & Tool Injection Surface — Are user-controlled inputs (chat messages, emails, documents fed into RAG) properly isolated from tool-calling logic? Can a malicious or cleverly crafted prompt cause the agent to exfiltrate data or perform unauthorized actions?

• Observability & Auditability — Is every prompt, tool call, argument, result and final decision logged in a tamper-evident way? Can you reconstruct exactly why an agent took a particular action six months later?

• Multi-Agent & Multi-Tenant Isolation — In swarm architectures (orchestrator + specialized agents), are there hard boundaries preventing one agent from escalating privileges or accessing another tenant's memory/RAG context?

• Supply Chain for Agents — Are you using third-party agent frameworks, custom runtimes or fine-tuned models? How do you verify they do not contain backdoors or unintended capabilities?

• Runtime Containment — Are agents executed in properly sandboxed environments (seccomp, AppArmor, gVisor, Firecracker, strict network policies) or do they run with broad host access 'because it was easier to develop'?

AI Trust Index gives development teams a structured, independent way to communicate the security posture of their agentic systems to clients, investors and regulators — moving beyond vague claims of 'we take security seriously'.

The New Era: From Antivirus to Anti-Agent Governance

We are witnessing a fundamental transition in defensive security. The previous generation focused on detecting and blocking malicious artifacts (files, processes, network indicators). The next generation must govern autonomous, goal-driven systems that were explicitly authorized to act on behalf of users and organizations.

This does not mean traditional defenses become irrelevant. It means they are now only the first layer. A second, equally critical layer is required: continuous trust assessment and behavioral governance of AI agents.

AI Trust Index is designed to be that second layer. It provides the missing 'credit score' for AI systems and the companies that build and operate them. Just as financial markets rely on independent credit ratings to allocate capital efficiently, the emerging agentic economy will need independent trust scores to allocate access to sensitive data and critical systems.

The organizations that understand this shift early — and can demonstrate high Trust Index scores for their AI initiatives — will have a decisive advantage in winning trust from customers, partners, regulators and investors.

Next Steps

To request an AI Trust Index assessment for your product, platform or internal AI systems, to integrate the scoring methodology into your development lifecycle, or to collaborate on the evolution of the standard:

- Email: kabzhanov@gmail.com
- Chat on the platform: bizdnai.com (bottom-right assistant)
- Open-source auditor tools: github.com/Kabzhanov/ati-audit
- Full platform: bizdnai.com — multi-agent business automation that already operates under AI Trust Index rules

We do not just publish scores. We help organizations build the next generation of trustworthy, governable agentic systems.

РУССКАЯ ВЕРСИЯ

Что такое AI Trust Index?

AI Trust Index — это независимый, вендоронейтральный стандарт и система скоринга, созданная специально для эпохи автономных ИИ-агентов и сложной автоматизации. Он измеряет, насколько хорошо организация или продукт понимает, управляет и коммуницирует риски, которые возникают, когда ИИ-системы получают реальную власть через tool-calling — доступ к файлам, базам данных, API, браузерам и мессенджерам.

В отличие от классических сертификатов безопасности, которые фокусируются на периметровых контролях или известных сигнатурах вредоносного ПО, AI Trust Index оценивает новые поверхности атак, появляющиеся при предоставлении интеллектуальным агентам легитимного доступа к бизнес-системам. Эти поверхности включают prompt injection, приводящий к непреднамеренным действиям, чрезмерные права инструментов (tool permissions), утечки памяти RAG между тенантами, цепочки привилегий в multi-agent системах и скрытый exfil данных через стандартные веб-протоколы, которые выглядят совершенно нормально для firewall и EDR.

Основные компоненты скора

- Учёт рисков — Систематически ли организация выявляет и документирует риски, специфичные именно для ИИ-агентов и автоматизации (а не только общие IT-риски)?
- Политика использования ИИ — Существуют ли чёткие, опубликованные правила, определяющие, когда, как и с какими правами ИИ-агенты могут работать с данными и системами компании?
- Юридическое и регуляторное соответствие — Соответствие ISO/IEC 42001 (системы менеджмента ИИ), Закону РК об ИИ (требования disclosure и consent), Закону о персональных данных 2026. Включает суб-скор G13 по готовности защиты персональных данных.
- Зрелость процессов — Есть ли repeatable-процессы для деплоя агентов, review прав, логирования каждого решения и tool call, а также incident response, специфичного для агентного поведения?
- Команда и governance — Есть ли в организации люди с явной ответственностью и компетенцией управлять автономными системами, или ИИ воспринимается как «просто ещё один инструмент»?
- Удовлетворённость пользователей и стейкхолдеров — Измеряется через прозрачную коммуникацию вовлечённости ИИ и реальное доверие пользователей к системам, с которыми они взаимодействуют ежедневно.

Privacy by Design: Все оценки проводятся в собственном окружении клиента. Публикуются только финальный скор, суб-скор G13 и высокоуровневое обоснование. Исходный код, сырые данные, внутренние документы и персональная информация никогда не покидают периметр клиента.

Почему классические средства защиты больше не достаточны

Firewall, VPN, антивирусы и даже современные EDR-системы создавались для обнаружения и блокировки вредоносного кода, пытающегося проникнуть извне или работать изнутри. ИИ-агенты не приходят в виде malware. Их явно приглашают через чат-интерфейсы,

API-интеграции, RPA-инструменты и кастомные скрипты автоматизации. Оказавшись внутри, они используют полностью легитимные протоколы (HTTPS, WebSocket, gRPC, OData, REST) и выглядят как обычные пользовательские или сервисные аккаунты.

Хорошо намеренный, но плохо изолированный ИИ-агент продаж, подключённый к WhatsApp и Bitrix24, может быть манипулирован через prompt injection и выгрузить всю базу клиентов. Интеграционный бот 1С с широкими правами OData может быть проинструктирован экспортировать данные склада на контролируемый атакующим эндпоинт — всё через «обычные» API-вызовы. Классические инструменты ничего не видят, потому что нет эксплойта, нет дропнутого файла, нет классического C2-бикона. Агент просто делает то, на что ему дали права.

Это ключевой сдвиг: мы перешли от «обнаруживать плохой код» к «управлять автономными агентами, которым дали легитимную власть». AI Trust Index — первый независимый механизм, который измеряет и коммуницирует, насколько хорошо это управление реализовано.

Для пользователей — Что нужно знать и требовать

Если вы используете ИИ-ассистентов, чат-автоматизацию или любой инструмент, который подключает интеллектуального агента к вашим рабочим данным (почта, CRM, 1С, документы, мессенджеры) — вы уже находитесь внутри новой threat model.

Риск не в том, что «кто-то взломает ваш компьютер». Риск в том, что ИИ-агент, которого пригласили вы или ваша компания, имеет больше доступа и меньше oversight, чем вы думаете. Один over-privileged агент может читать, суммировать и передавать конфиденциальную информацию через каналы, которые выглядят как нормальная работа.

Практические вопросы, которые каждый пользователь должен задавать вендорам и внутреннему IT:

- Есть ли у этого ИИ-инструмента или агента независимый AI Trust Index score? Могу ли я увидеть детальное обоснование?
- Что именно этот агент может читать, изменять или отправлять? Есть ли чёткий, аудитируемый список его tool permissions?
- Куда уходят мои данные, когда агент их «обрабатывает» или «суммирует»? Логируется ли это и могу ли я аудитить каждое решение?
- Если агента скомпрометируют или манипулируют (например, через умело составленный промпт), каков радиус поражения?
- Публикует ли вендор, как он изолирует агентов разных клиентов и предотвращает cross-tenant утечки данных?

Вы имеете право требовать прозрачности. AI Trust Index существует именно для того, чтобы пользователи могли принимать информированные решения, а не слепо доверять фразе «у нас продвинутая ИИ-безопасность».

Для компаний — Бизнес-, регуляторные и репутационные риски

Для любой организации, которая развертывает ИИ-агентов внутри или предлагает ИИ-автоматизацию клиентам, вопрос уже не «есть ли у нас антивирус и firewall?». Вопрос: «как мы докажем, что наши автономные системы достойны доверия и находятся под контролем?».

Новый Закон РК об ИИ и Закон о персональных данных (2026) вводят явные требования по disclosure взаимодействий с ИИ и получению согласия. Несоответствие — не теоретическая угроза, а реальная юридическая экспозиция. Помимо регуляторики, один громкий инцидент с утечкой данных через ИИ-агента может уничтожить годы доверия клиентов и партнёров.

Ключевые бизнес-импликации:

• Управление рисками вендоров — Перед подключением любого стороннего ИИ-агента, бота автоматизации или интеграции (особенно касающихся 1С, Bitrix24, amoCRM, WhatsApp или складов) требуйте от вендора публикации актуального AI Trust Index score и отчёта.

• Внутренние ИИ-инициативы — Кастомные агенты, разработанные вашими командами или подрядчиками, должны проходить оценку. «В тесте работает» недостаточно, когда агент имеет доступ к production-данным.

• Отчётность инвесторам и совету директоров — AI Trust Index даёт единый, сопоставимый, защищаемый метрику, которую можно включать в отчёты по security posture, due diligence пакеты и ESG-дисклозуры.

• Конкурентное преимущество — Компании, которые могут продемонстрировать высокие Trust Index scores для своих ИИ-систем, получают преимущество при тендерах с клиентами, чувствительными к безопасности, и в регулируемых отраслях.

• Готовность к инцидентам — Организации со зрелым agent governance (логирование решений, чёткие границы прав, быстрые процедуры containment) быстрее восстанавливаются и несут меньшие регуляторные последствия при инцидентах.

Для разработчиков и архитекторов — Новая threat model

Если вы разрабатываете, интегрируете или эксплуатируете ИИ-агентов, multi-agent системы, RAG-пайплайны, боты автоматизации или любые LLM-оркестрированные workflow, которые касаются реальных бизнес-данных или систем — парадигма безопасности кардинально изменилась.

Классическая модель application security (валидация ввода, output encoding, least-privilege аккаунты в БД) необходима, но уже недостаточна. Агенты вводят новые классы рисков, потому что они goal-oriented и tool-augmented. Их можно манипулировать не только через уязвимости кода, но и через инструкции на естественном языке, которые меняют поведение в runtime.

Критические области, которые AI Trust Index оценивает у команд разработки:

• Excessive Agency (чрезмерная агентность) — Есть ли у агента больше инструментов и прав, чем строго необходимо для заявленной цели? Может ли он читать всю файловую систему, выполнять произвольные команды или обращаться к production-БД, когда ему нужно только отвечать на вопросы по одному модулю?

• Поверхность prompt & tool injection — Правильно ли изолированы контролируемые пользователем вводы (сообщения в чате, письма, документы в RAG) от логики tool-calling? Может ли вредоносный или умело составленный промпт заставить агента exfiltrate данные или выполнить несанкционированные действия?

• Observability & Auditability — Логируется ли каждый промпт, tool call, аргумент, результат и финальное решение tamper-evident способом? Можно ли через полгода точно реконструировать, почему агент принял то или иное решение?

• Изоляция в multi-agent и multi-tenant — В архитектурах роя (оркестратор + специализированные агенты) существуют ли жёсткие границы, предотвращающие эскалацию

привилегий одного агента или доступ к памяти/RAG-контексту другого тенанта?

• Supply chain для агентов — Используете ли вы сторонние agent frameworks, кастомные runtime или fine-tuned модели? Как вы верифицируете, что в них нет backdoors или unintended capabilities?

• Runtime containment — Выполняются ли агенты в правильно sandboxed окружениях (seccomp, AppArmor, gVisor, Firecracker, строгие network policies) или работают с широким доступом к хосту «потому что так было проще разрабатывать»?

AI Trust Index даёт командам разработки структурированный, независимый способ коммуницировать security posture своих агентных систем клиентам, инвесторам и регуляторам — выходя за рамки расплывчатых заявлений «мы серьёзно относимся к безопасности».

Новая эра: от антивируса к governance автономных агентов

Мы наблюдаем фундаментальный переход в defensive security. Предыдущее поколение фокусировалось на обнаружении и блокировке вредоносных артефактов (файлы, процессы, сетевые индикаторы). Следующее поколение должно управлять автономными, goal-driven системами, которым явно дали полномочия действовать от имени пользователей и организаций.

Это не означает, что традиционные средства защиты теряют актуальность. Это означает, что они теперь только первый слой. Требуется второй, не менее критичный слой: непрерывная оценка доверия и поведенческое управление ИИ-агентами.

AI Trust Index создан именно как этот второй слой. Он предоставляет недостающий «кредитный скоринг» для ИИ-систем и компаний, которые их создают и эксплуатируют. Как финансовые рынки полагаются на независимые кредитные рейтинги для эффективного распределения капитала, так и формирующаяся агентная экономика будет нуждаться в независимых скорах доверия для предоставления доступа к чувствительным данным и критичным системам.

Организации, которые поймут этот сдвиг первыми — и смогут демонстрировать высокие Trust Index scores для своих ИИ-инициатив — получат решающее преимущество в завоевании доверия клиентов, партнёров, регуляторов и инвесторов.

Следующие шаги

Чтобы запросить оценку AI Trust Index для вашего продукта, платформы или внутренних ИИ-систем, интегрировать методологию скоринга в жизненный цикл разработки или поучаствовать в развитии стандарта:

• Email: kabzhanov@gmail.com
• Чат на платформе: bizdnai.com (ассистент в правом нижнем углу)
• Open-source инструменты аудитора: github.com/Kabzhanov/ati-audit
• Полная платформа: bizdnai.com — multi-agent автоматизация бизнеса, которая уже работает по правилам AI Trust Index

Мы не просто публикуем скоры. Мы помогаем организациям строить следующее поколение достойных доверия и управляемых агентных систем.