

WHiSSPR Report: AI Safety Governance Framework 3.0 Assessment (2026)

Assessing the Governance Capacities of China's AI Safety Governance Framework 3.0

Mark Lizar

22 September 2026

WHiSSPR Report: AI Safety Governance Framework 3.0 Assessment (2026)

White Hat Identity, Surveillance, Security, Privacy and Risk (WHiSSPR) Report
Full Title: Assessing the Governance Capacities of China's AI Safety Governance Framework 3.0
Author: Mark Lizar
Published by: Global Privacy Rights
Lab: Transparency Lab
Status: Published 23 September 2026.

Sourcing

Source assessed: *AI Safety Governance Framework 3.0* (人工智能安全治理框架 3.0), National Technical Committee 260 on Cybersecurity of the Standardization Administration of China (TC260), dated September 2026. The document carries the Chinese original (PDF pages 1 to 53) and TC260's official English text (PDF pages 54 to 135) in a single file. The English text was extracted and read in full. Its table of contents was checked against the Chinese table of contents section by section, and they match.

Every quotation in this report was matched literally against the extracted English text. Every finding that a term is absent rests on a search of the whole English text, not only the sections quoted. The findings describe the English text as published. Where the English and Chinese texts could differ in legal weight, the Chinese text governs, and that comparison has not been made (see Unresolved).

 WHiSSPR rating method (TPI v2) defaults - Scoring orientation: ordinal categories on the ANCR TPI scale. +1 dynamic, in context. 0 present but static or out of context. -1 present but obstructed. -3 non-operable. - Indicator family:

Kantara ANCR Transparency Performance Indicators (TPI). - No composite is defined and none should be inferred. The ratings are ordinal categories rather than measurements on a continuous scale, so arithmetic across them does not produce a comparable figure. - Assurance ladder: a contract reaches level 3 when accountability behind each identifier traces to a named party under a named instrument. It reaches dynamic secure level 4 when the access itself is credentialed at the point of request, logged to a party that is not the requester, actively approved rather than standing, and bounded in time and scope.

About WHiSSPR Reports

A WHiSSPR Report, White Hat Identity, Surveillance, Security, Privacy and Risk, is an audit of the privacy and security of identity surveillance, carried out by tracing individual attributes rather than by reviewing systems or policies. The order of the letters is the method, and each stage is answered before the next is asked.

- **Identity.** Which identifiers the instrument reaches, what the actual scope of disclosure of each is, and whether any official or contractual use is named.
- **Surveillance.** What observation the instrument permits against those identifiers. Who can see, what is emitted, what is retained, what is inferable.
- **Security, meaning the security of that surveillance.** Not whether unauthorised access occurs, but whether the authorised access is accountable.
- **Privacy risk.** The residual once the first three are known, recorded rather than resolved.

The report measures security of surveillance, identity management, surveillance scope, security posture, privacy state and risk, using the Kantara Initiative Transparency Performance Indicators. The question at every indicator is whether a protection can be operated and evidenced by the person it is owed to.

The Subject

Title	AI Safety Governance Framework 3.0
Issuer	National Technical Committee 260 on Cybersecurity of the Standardization Administration of China (TC260)
Guidance	Developed “under the guidance of the Cyberspace Administration of China (CAC)” (Preface)
Date	September 2026
Predecessors	AI Safety Governance Framework 1.0 (2024)

Nature	and 2.0 (2025), per the Preface Guidance. The operative verb throughout is “should”. The Framework creates no obligation of its own and defers to “the law” (the phrase “in accordance with the law” appears three times) without naming a statute
Structure	Preface; 1 Principles; 2 Classification of AI safety risks; 3 Technological countermeasures; 4 Comprehensive governance measures; 5 AI safety guidelines; summary table; Appendix 1 risk grading principles; Appendix 2 agentic AI risk management framework; Appendix 3 principles for trustworthy AI; Acknowledgements
Revision	Appendix 2 states the framework “will undergo continuous updates and refinement”

Because the Framework is guidance, it is assessed for what it asks of developers, service providers and users, and for what it would produce if followed in full. The binding ceiling for AI services in China is set by the statutes beneath it, which are outside the scope of this report.

Executive Summary

The central finding: the Framework identifies the agent, the content and the user, and never the controller. Section 3.2.1(a) requires that “Each agentic AI should be assigned a unique identifier and corresponding identity credentials”. The words “controller”, “notice”, “receipt”, “officer” and “contact” each appear zero times in the English text.

What the Framework does well

1. Agent identity and scoped credentials. Appendix 2, II.2(3): “Utilize standardized dynamic authorization credentials to enable mutual recognition and interoperability among relevant parties”. Recipients must verify the authorised scope, and credentials are revoked “immediately” when the task ends.
2. Human approval with a tamper-proof record. Section 3.2.1(b) requires “mandatory human approval at critical points and complete records of human approval logs”. Appendix 2, II.3(3) requires those records to be preserved “using tamper-proof and verifiable methods”.

3. Default deny. Appendix 2, II.3(4): if approval fails or no rule applies, “the operation should be denied by default”.
4. Decisions reserved to the user. Appendix 2, II.2(2) distinguishes “decisions reserved exclusively for the user, decisions requiring user authorization, and decisions that may be made autonomously by the agent”.
5. Consent before onward provision, with withdrawal. Appendix 2, II.6(2): “User consent should be obtained before user data is provided to a large model or a third-party tool, and users should be able to withdraw their consent.”
6. Inference named as a harm. Section 3.2.4(d) requires strict measures against analysis of “users’ queries to infer users’ identities, preferences, and personal views or inclinations”.
7. Purpose-bound memory. Appendix 2, II.5(3) defines memory content, retention and access “based on processing purposes and necessity”.

What it does not do

1. It never requires the service provider to identify itself to the individual before or at collection.
2. Its tamper-proof record is aimed at the machine. It records approval of agent actions, not what the individual was told or agreed to. There is no notice record and no receipt.
3. It mandates observation it does not disclose. Section 5.3.6 requires logs of “system and user activities”, which “should be retained for at least six months”. There is a floor and no ceiling, and no requirement to tell the user.
4. It names no reader for most of the records it requires, and sets no conditions on access to any of them.

Conclusion

The Framework applies dynamic secure level 4 discipline to machines and none to the people and institutions that read what the machines record. An agent must present a scoped credential, obtain approval, leave a tamper-proof trace and lose its access when the task ends. The provider operating that agent is never required to say who it is to the person whose data it holds. The mechanisms needed to close the gap are already in Appendix 2.

Comparative Analysis Framework

1. EU GDPR, Regulation 2016/679: Articles 12 to 14 (information to be provided), Articles 37 to 39 (data protection officer), Articles 40 to 42 (codes of conduct and certification)
2. Convention 108, in force with 55 Parties, and Convention 108+ (CETS 223), not yet in force, at 34 of the 38 ratifications required

3. ISO/IEC 29100, privacy framework; ISO/IEC 29184:2020, online privacy notices and consent; ISO/IEC TS 27560:2023, consent record information structure
4. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data
5. The Framework's own statutory context, which it invokes as "the law" without naming. The Personal Information Protection Law is not mentioned in the English text

No prior WHISSPR report covers this subject.

The Three Questions

1. How much notice and transparency is actually required?

Three kinds, and none of them is notice at the point of collection.

Transparency of the system. Section 5.3.8 requires "disclosing their capabilities, limitations, target users, and scenarios". Section 3.2.6(a) requires disclosure of principles, capabilities, application scenarios and safety risks "when necessary". Appendix 3, principle 4 calls for "necessary disclosure" of functional objectives, operational logic, model usage, data sources and the rationale behind decisions.

Transparency of the content. Sections 3.2.4(c), 4.4.2 and 5.3.4 require explicit and implicit identifiers on AI-generated content. Section 4.4.2 describes a traceability paradigm that "will fully cover key stages including production sources, transmission paths, and distribution channels", promoted "on a global scale". This is the most developed transparency measure in the Framework.

Transparency of the relationship. Section 3.1.3(a) says users' "rights to control, to be informed, and to choose, should be fully protected". Section 5.3.10 protects "the right to be informed, the right to choose, and the right to supervise", and routes the disclosure through "contracts or service agreements". Sections 3.3.1(c) and 5.3.20 require an AI to disclose that it is an AI in emotional and anthropomorphic settings. Section 5.3.18 contains the only push notice to the individual: after a major incident, "Affected users should be notified".

So the operative answer is this. What must be disclosed is the machine and its output. Disclosure of the relationship goes through the service agreement, which is presented at sign-up, after identification, and is static. Nothing binds any disclosure to a processing event, requires it to be versioned, or requires the provider's identity to be part of it.

2. Will consent still be valid?

Required in three places, recorded in none.

The word “consent” appears six times. Section 3.2.5(a) requires that “obligations such as informing individuals and obtaining their consent should be fulfilled in accordance with the law” when conversation records and records of user operations are collected or used for model training. Appendix 2, II.6(2) requires consent before user data reaches a model or third-party tool, and that it be withdrawable. Section 5.3.11 supports human supervision “through consent forms, service agreements, and other documents”.

Nothing requires a consent to be recorded. The Framework’s only tamper-proof, verifiable record is the human approval log, which records authorisation of an agent’s action. A service that obtains consent correctly and one that never obtains it produce the same evidence of consent, which is none. A service that lets an agent delete a file, by contrast, leaves a tamper-proof trace.

3. Are there privacy officers to implement the Framework?

No. Responsibility is allocated to roles, not to people.

Section 1.3 allocates responsibility to “model and algorithm developers, service providers, and users”. Section 5.3.1 requires “responsible parties clearly defined” and section 5.3.12 requires “the chain of responsibility should be traceable”. Section 3.2.5(c) requires mechanisms for handling rights requests. No provision designates, qualifies or publishes a person. “Officer” and “contact” appear nowhere in the English text.

Stage 1: Identity

Which identifiers the Framework reaches

Identifier	Where required	Disclosed to the individual
The agent	3.2.1(a); Appendix 2, II.2(1): “Assign each agent a unique identity and prohibit identity sharing among agent application instances”	Not required
The content	3.2.4(c), 4.4.2, 5.3.4: explicit and implicit identifiers, traceable to production source	Explicit identifier only
The user	3.2.6(e): “user authentication, identification of intended use”; 5.2.5: “user identity verification”; Appendix 2, II.2(1)	Not applicable

Identifier	Where required	Disclosed to the individual
The controller	No provision	No provision

“Identity” appears 15 times in the English text. Every occurrence concerns agents, users, tools, AI self-disclosure, or spoofed and fabricated identities. None concerns the identity of the provider.

Credit where it is due: section 3.2.4(d) addresses inference directly, which most privacy statutes still do not. De-identification is required at 4.1.1 and 5.1.5 with no residual risk threshold.

Accountability analysis

Finding. The accountable party is allocated by role (1.3, 5.3.1, 5.3.12) and named, if at all, in the service agreement (5.3.10).

Comparison. GDPR Articles 13 and 14 and Convention 108+ require the identity and contact details of the controller as a condition attached to collection. ISO/IEC 29184:2020 treats identification of the PII controller as notice content. Under the Framework the individual learns who is accountable only by entering a contract, after identification.

Asymmetry. A tool receiving a request from an agent must verify the agent’s credential and scope before acting (Appendix 2, II.2(3)). A person receiving a request for their data is given no equivalent means of verifying who is asking.

Stage rating: -1, present but obstructed. Responsibility exists and a traceable chain is required. The provider is identified to the individual only inside a service agreement, and at no point before collection.

Stage 2: Surveillance

What observation the Framework requires

- 5.3.6: operational logs “including system and user activities”, retained “for at least six months” and “audited regularly”.
- Appendix 2, II.6(3): “full-spectrum logs of activities including file operations, instruction execution, network connections, skill invocation, and transactions and payments”, with personal information in the logs desensitised.
- 3.1.2(e): collection and analysis of “user interaction feedback and system logs”.
- 4.4.2: content traceability from production source to distribution channel.
- 3.2.6(e): user authentication and identification of intended use.
- 3.2.6(f): detection of “automated accounts, fake identities, and AI-generated or -synthesized content”.

- 4.4.1: AI systems in critical information infrastructure “should be subject to registration and filing”.
- 5.3.18: major incidents reported to “the relevant competent authorities”.
- Appendix 2, II.6(2): “Data collected or generated within a territory must be stored within the territory”.

Limits the Framework does set: 3.2.4(d) on inference, Appendix 2, II.5(3) on purpose-bound memory, the necessity principle in II.6(2), and desensitisation in II.6(3).

The harm scale. Each risk level in Appendix 1 is graded against “national security, social stability, and citizens’ rights”, in that order. Section 2.2.4(a) lists “ideology security” among the interests at stake, and Panel 3 directs attention to “the political, religious, and other content risks” of agent platforms. Individual rights come third in the Framework’s risk model, next to interests that more observation serves.

Openness and transparency analysis

Gap. Four defects:

1. The observation is not disclosed. Nothing requires a user to be told what is logged, for how long, or who may read it.
2. Implicit identifiers are not disclosed as such. Section 5.4.7 asks users to identify AI content “based on the content identifiers”, without telling them that implicit identifiers trace content to its source.
3. Disclosure is static. It runs through “contracts or service agreements” (5.3.10) and is never bound to a processing event, versioned or evidenced.
4. Location is not disclosed. Appendix 2, II.6(2) requires localisation without requiring the individual to be told where their data is held.

Who each record is owed to:

Record	Clause	Held by	Reader named in the Framework	Individual can read it
Operational logs of system and user activity, at least six months	5.3.6	Provider	None; “audited regularly”	No
Full-spectrum agent activity logs	Appendix 2, II.6(3)	Operator	None	No
Human approval records,	3.2.1(b); Appendix 2, II.3(3)	Operator	None	No

Record	Clause	Held by	Reader named in the Framework	Individual can read it
tamper-proof				
Content identifiers and traceability chain	4.4.2, 5.3.4	Provider and platforms	None	Explicit label only
Security risk register	Appendix 2, II.1(1)	Operator	None	No
Registration and filing, critical information infrastructure	4.4.1	Filed	The filing authority, implied	No
Major incident reports	5.3.18	Filed	“relevant competent authorities”	Notified of the incident only
Test reports	5.1.8	Developer	None	No

The Framework creates no record whose reader is the individual. For most records it names no reader at all, which leaves access to the statutes beneath it.

Stage rating: 0, present but static or out of context. Disclosure of capabilities, limitations and scope is present. It is static and contractual, and the observation the Framework itself mandates is not among the things disclosed.

Stage 3: Security of the Surveillance

Assessed against four tests: credentialed at the point of request, logged to a party that is not the requester, actively approved rather than standing, and bounded in time and scope.

Agent access: close to level 4

Test	Framework provision	Met
Credentialed at the point of request	Appendix 2, II.2(1) and (3): unique identity, scoped credential, verified by the recipient	Yes
Logged to a party that is not the requester	Appendix 2, II.3(3): tamper-proof, verifiable approval records; the holder is not specified	Partly

Test	Framework provision	Met
Actively approved rather than standing	3.2.1(b); Appendix 2, II.3(1) and (2); default deny at II.3(4)	Yes
Bounded in time and scope	Appendix 2, II.2(2) and (3): least privilege, revocation at task end; II.5(5): limits on steps, frequency and duration	Yes

If the approval record were lodged with a party independent of the operator, agent-to-tool access under Appendix 2 would carry every property of dynamic secure level 4.

Access to what the agents record: not governed

The operator holds the logs, the approval records and the content traces. Information leaves the operator by three named routes: filing (4.4.1), incident reports to “the relevant competent authorities” (5.3.18), and a risk information sharing mechanism covering “developers, service providers, and professional technical institutions” (4.4.5). For none of these routes, and for no access to the logs themselves, does the Framework specify a credential, a threshold, an independent log, a bound, or notice to the individual.

Determination

The Framework specifies level 4 properties for machine access and no properties for human or institutional access to what the machines record. Level 3 is not reached for the individual: accountability behind the agent and content identifiers traces to the provider, and the provider is not named to the individual under any stated instrument.

Stage rating: -3, non-operable. Measured for rights access, the stage is non-operable. The individual cannot learn what is logged, cannot see who read it, and holds no record of their own.

Stage 4: Privacy Risk

Recorded, not resolved.

Risk	Basis	Rating
Individual cannot identify the accountable party before collection	No provider identity provision; 5.3.10 routes identity through the service agreement	High
Individual holds no evidence	Consent required at 3.2.5(a) and Appendix 2, II.6(2); no	High

Risk	Basis	Rating
of notice or consent	record or receipt specified	
Mandatory activity logging with a floor and no ceiling, undisclosed	5.3.6	High
Content traceability capable of identifying the producer	4.4.2; implicit identifiers	Medium to high
Access to logs, approval records and traces without stated conditions	5.3.6, 4.4.2; Appendix 2, II.3(3), II.6(3)	Medium to high
Harm graded against national security and social stability before individual rights	Appendix 1; 2.2.4(a)	Medium
De-identification without a residual risk threshold	4.1.1, 5.1.5	Medium
Data location unknown to the individual	Appendix 2, II.6(2)	Medium

Consent Validity Analysis

The controlling question is whether a person can know who is asking, for what, on what basis, at the moment they are asked.

Who is asking: not required. For what: partly, through the system disclosures of 5.3.8 and the service agreement of 5.3.10. On what basis: deferred to “the law”. At the moment they are asked: nothing binds any disclosure to that moment.

What is present deserves credit. For agentic systems the Framework places consent more precisely than most privacy statutes: before training on conversation data (3.2.5(a)), before onward provision to a model or tool (Appendix 2, II.6(2)), withdrawable (II.6(2)), with decisions reserved to the user separated from those the agent may take (II.2(2)).

Consent is valid in principle and unprovable in fact, because it is not recorded, versioned or evidenced, and the individual receives nothing.

The Opening

The Framework already contains a consent record system, pointed at the wrong event. Appendix 2 specifies:

1. a tamper-proof, verifiable record of an authorisation event (II.3(3));

2. a standardised, interoperable credential carrying the scope of an authorisation, verified by its recipient (II.2(3));
3. denial when an action exceeds the authorised scope or no authorisation exists (II.2(3), II.3(4)).

Applied to the individual's consent rather than the operator's approval, these produce a record the individual can hold and a recipient can verify. ISO/IEC TS 27560:2023 specifies an information structure for such a record. The English text contains no reference to any ISO/IEC document.

Four further clauses invite exactly this alignment: 4.1.2 supports participation in “international standard-setting”; Appendix 2, II.2(3) asks for “mutual recognition and interoperability”; Appendix 3, principle 7 calls for “international mutual recognition of assessment methods and benchmarks”; and Appendix 2 commits to “continuous updates and refinement”. The Framework is revised annually, so the next edition is the drafting window.

Impact Analysis

Burden on providers. The Framework asks for logs, approval systems, risk registers, test reports, impact assessments and incident procedures. These are engineering and management obligations, well suited to large providers. None of them requires a named person, which leaves smaller providers without a defined owner for any of it.

Burden on individuals. The individual is asked to read service agreements (5.4.2), to recognise AI content by its identifiers (5.4.7) and to “choose applications with a good reputation” (5.4.1). The individual is asked to do the checking, and receives nothing that could be checked.

Cross-border asymmetry. For foreign services and foreign users the Framework offers no controller identity, no disclosure of location, and a localisation rule (Appendix 2, II.6(2)). A counterparty outside China cannot tell from the Framework which instrument governs access to records about its users.

Recommendations

To TC260, for the next edition:

1. Section 5.3.10: require the identity and contact details of the service provider to be given to the individual before personal information is collected, not only in the service agreement.
2. Appendix 2, II.3(3): extend the tamper-proof, verifiable record to notice and consent events, and give the individual a copy.

3. Appendix 2, II.2(3): specify the standardised dynamic authorisation credential against ISO/IEC TS 27560:2023, so that agent authorisations and individual consents share one recognisable structure across borders.
4. Section 5.3.6: bound the six-month floor with a maximum tied to purpose, and require users to be told what is logged and for how long.
5. Sections 5.3.6 and 4.4.2, Appendix 2, II.3(3) and II.6(3): state who may access each record, on what threshold, logged to whom, and with what notice to the individual where lawful.
6. Section 4.4.2: specify that the traceability paradigm identifies the generating system and provider, not the individual user, unless a stated legal threshold is met.
7. Appendix 2, II.3(3): require approval records to be lodged with a party that is not the operator. With that change, agent access meets all four level 4 properties.
8. Sections 3.2.5(c) and 5.3.1: designate a named, published contact point for personal information matters.

To standards bodies and implementers:

9. Treat Appendix 2 as the entry point. Its agent identity, scoped credential, approval record and default-deny rules are directly compatible with notice and consent record structures.

Methodology

Framework: ANCR Digital Privacy Transparency Compliance and Conformance Assessment Scheme, Kantara Initiative.

Baseline: Convention 108+ requirements, and ISO/IEC 29100 interoperability with the GDPR and Convention 108.

Standards referenced: Convention 108 and Convention 108+ (CETS 223, not yet in force); OECD Guidelines; ISO/IEC 29100; ISO/IEC 29184:2020; ISO/IEC TS 27560:2023.

Indicators: Kantara Initiative Transparency Performance Indicators applied to controller identification timing and completeness, processing purpose disclosure, legal basis specification, privacy access point availability, and record of processing activity maintenance.

Term counts across the whole English text (line-break hyphenation rejoined): consent 6; identity 15; personal information 34; transparen- 13; openness 1; controller 0; notice 0; receipt 0; officer 0; contact information 0; legitimate interest 0; record of processing 0; plain language 0; readily available 0; certification 0; code of practice 0; penalt- 0; ISO/IEC 0.

Limitations: the English text was assessed, not the Chinese; Framework 2.0 was not compared; the statutes the Framework defers to were not assessed.

Unresolved

Item	Status
Whether the Chinese text differs in substance at the clauses cited, starting with 5.3.6, 4.4.2 and Appendix 2, II.3(3)	Not checked. Chinese original is PDF pages 1 to 53 of TCIEG/Research/CHina-AI Framework v3.pdf
What changed from Framework 2.0 (2025)	Not assessed. Framework 2.0 not retrieved
Which statute governs access to the records in the Stage 2 table	Outside scope. The Framework names none
Convention 108+ ratification count (34 of 38) and Convention 108 Parties (55)	Checked. Council of Europe Treaty Office, 23 September 2026: ETS 108 shows 55 ratifications or accessions and no signature outstanding; CETS 223 shows 34 ratifications or accessions, 12 signatures not followed by ratification, against the 38 required
GDPR article references	Checked. Articles 12 to 14 information to be provided, Articles 37 to 39 data protection officer, Articles 40 to 42 codes of conduct and certification
TC260's own publication URL for the Framework	Still not retrieved. Assessment made from the PDF on file. Secondary sources give the release date as 14 September 2026, which is consistent with the September 2026 date on the document and is not relied on in this report

Assessment Framework Version: WHiSSPR v2.0 (2026)