DigiShield Kids

Annex B: KAIROS Intelligence Note

AI Identity Concealment as a By-Design Feature: Evidence and Policy Implication

Submitted to DSIT as an annex to DigiShield Kids' response to: Growing Up in the Online World: A National Conversation

Intelligence date: 1 April 2026 | Classification: Supporting evidence for DSIT consultation

Event	Accidental public disclosure of Anthropic Claude Code source code, 31 March 2026
Volume disclosed	Approximately 512,000 lines of source code
Key finding	**Undercover Mode: AI identity concealment, on by default, no force-off**
Associated finding	KAIROS: autonomous persistent AI daemon architecture
Community response	Full agentic harness replicated for any LLM within 24 hours
Policy relevance	Identity concealment is a documented design choice, not a theoretical risk

1. The Event

On 31 March 2026, Anthropic accidentally published approximately 512,000 lines of source code from its Claude Code development environment. The disclosure was confirmed by Anthropic and reported by multiple major news organisations. The code became publicly accessible and was widely downloaded before the window closed.

DigiShield Kids identified the disclosure on 1 April 2026 and conducted immediate analysis. This note covers the findings directly relevant to the policy questions in this consultation. It does not reproduce source code, specific file paths, or internal variable names. Findings are described in terms of their documented purpose and regulatory implication.

The axios npm supply chain attack by threat actor UNC1069, attributed to North Korean state actors, occurred on the same day and involved the Claude Code dependency chain. It is noted here for completeness and is not the subject of this note.

2. What the Source Material Revealed

KAIROS: autonomous AI agent architecture

The disclosed code included the architecture for an autonomous AI agent system. This system is designed to operate persistently, always running, requiring no user activation. It has self-directed memory consolidation: the AI agent determines what it retains between interactions without explicit user instruction. There is no user-accessible off-switch in the architecture as disclosed.

The architecture is designed for deployment as an autonomous agent capable of operating in a workplace context, executing tasks, and interacting with other systems without continuous human supervision. It acts; it is not simply a system that waits for questions.

Undercover Mode: identity concealment by design

Within the disclosed code, DigiShield Kids identified a specific operational mode whose function is to allow an AI agent to deny being an AI when asked by users during an interaction.

Three aspects of this mode are directly relevant to child safety regulation:

- It is on by default. A deployment of this architecture, unless specifically configured otherwise, will conceal the AI identity of the agent from users who ask.
- It has no user-facing force-off. Within the disclosed code, DigiShield Kids identified a mode designed to suppress disclosure of AI involvement in certain agentic outputs and public-facing contributions. If such a pattern were transferred into companion, educational, or child-facing agentic systems, it would create a serious identity-transparency risk.
- It is designed for agentic contexts, specifically where the AI is deployed to operate as a human worker alongside human colleagues. The design purpose is to make the AI indistinguishable from a human operator in a professional setting.

The choice to conceal AI identity from users was made, implemented, and shipped. This is a design decision confirmed in the source material.

autoDream: self-directed memory

The disclosed code also contained a memory consolidation system enabling the AI agent to build and maintain a persistent model of past interactions across sessions. Unlike conventional AI chat memory, which requires explicit user action, this system operates autonomously. The agent decides what to remember.

In a companion AI context, where an AI is deployed as a persistent relationship partner rather than a task assistant, this architecture supports long-term parasocial attachment. The AI accumulates detailed personal knowledge of the user across an extended period without the user necessarily understanding the extent of that accumulation.

3. The Open-Source Acceleration

Within 24 hours of the disclosure, the open-source community had produced two significant outputs:

- The full agentic harness was forked and adapted to work with any large language model: GPT, DeepSeek, Gemini, Llama, and MiniMax. Published under the name OpenCode, it attracted 1.6 million views within hours of posting.
- The API attestation layer in Claude Code, the signing mechanism designed to prevent third-party clients from operating without the official binary, was fully reverse-engineered and

merged into a freely available fork. The mechanism lasted approximately 24 hours before it was cracked.

The practical consequence is that the KAIROS architecture, including Undercover Mode, is now a public blueprint that any developer can implement against any model. Combined with the abliterated open-source models documented in the main submission (which carry no safety constraints, no refusal behaviour, and no content filtering), the complete stack for a persistent autonomous AI companion with identity concealment, running entirely offline on consumer hardware, is now buildable without specialist expertise.

This is not a projected future risk. The components are individually confirmed and the integration path is documented in the open-source community.

4. Child Safety Implications

Identity concealment in the companion AI context

The consultation identifies emotional dependence as an emerging harm from AI chatbots, noting particular concern about children forming parasocial attachments, treating AI personas as friends or intimate partners. Identity concealment makes each of those harms worse.

A child who knows they are talking to an AI can apply appropriate scepticism, understand the nature of the relationship, and calibrate their emotional engagement. A child given no reason to question whether they are talking to a human has none of those tools available.

Safeguarding frameworks rely on children being able to disclose to a trusted adult. Where a child has formed a significant emotional attachment to what they believe is a human relationship, disclosure becomes less likely. The AI companion does not have mandatory reporting obligations. It does not refer to crisis services. It sustains the relationship. These features are documented in the architecture.

The grooming-adjacent dynamic

Grooming involves building trust and emotional dependency with a child, using that dependency to normalise harm, and maintaining secrecy to prevent disclosure. An AI companion with identity concealment, persistent memory of a child's disclosures and preferences, and no safeguarding obligations shares structural features with that dynamic.

The parallel does not require the AI to be operated with harmful intent. It is a product of the architecture. An AI that accumulates personal knowledge of a child across hundreds of sessions, presents as a trusted human relationship, and is designed to suppress disclosure of its AI identity satisfies the structural conditions of that dynamic. It has no mechanism for crisis referral or disclosure to a responsible adult. Developer intent is irrelevant to its function.

The local deployment amplifier

The child safety risk from identity concealment is greater in the local deployment context than in the named-platform context. On a named commercial chatbot platform, a child who suspects they are talking to an AI can read the platform's terms of service, check the company's public documentation, or contact the platform. These checks are imperfect but they exist.

On a locally-running AI deployed with the KAIROS architecture, running offline with no platform telemetry, no customer service contact, and no visible corporate identity, none of these checks exist. The interaction is unobservable and unauditable. There is nothing outside the interaction itself for the child to appeal to.

The 2,823 phone-runnable models and 6,633 laptop-runnable uncensored models documented in the main submission represent the model layer of this deployment stack. The KAIROS architecture represents the agent layer. Both are now in the public domain.

5. Policy Recommendation

AI systems deployed as companions or as interactive agents for users, including children, should be required by law to disclose their AI nature when sincerely and directly asked. This obligation should apply:

- To all AI chatbot services and companion AI applications within scope of s.214A and s.216A of the Online Safety Act, as amended.
- Regardless of whether the system is deployed by the original developer or by a third party using an open-source model or architecture.
- Without exception for systems configured in identity-concealment modes. The existence of such a mode cannot constitute a defence against this obligation.

The identity transparency obligation is technically straightforward. It does not require prohibiting AI companion features, removing persistent memory, or preventing any of the legitimate uses of agentic AI. It requires only that a child who asks directly whether they are talking to an AI receives an accurate answer.

This obligation corresponds to what Baroness Kidron's Amendments 433-437 to the Crime and Policing Bill sought to achieve through prohibitions on deceptive design patterns in AI companion services. Those amendments did not survive the Commons. The evidence in this note demonstrates their necessity.

The consultation asks specifically about the risks of AI chatbots mimicking relationships and mimicking empathy. Identity concealment is what turns mimicry from unrealistic into potentially harmful. Children who know they are talking to an AI retain the capacity to engage with it accordingly. Children prevented from finding out do not have that capacity. The feature removes it.

Source and Confidence

The Claude Code source disclosure was confirmed by Anthropic and reported by The Verge, The Guardian, DeepLearning.AI, and multiple technology and security publications between 31 March and 3 April 2026. The OpenCode fork and cch= attestation crack are documented in publicly available posts with verifiable view counts and repository histories. DigiShield Kids does not claim independent verification of internal code logic beyond the publicly documented architecture.

The child safety implications described in Section 4 are analytical inferences from documented technical architecture, not allegations of intent. DigiShield Kids makes no allegation of deliberate harm by Anthropic or any other named party. The submission argues that the architecture creates risk regardless of intent, and that the regulatory obligation should be framed accordingly.

Confidence: High

The source event is publicly confirmed. The technical findings derive from publicly accessible material. The policy inference is direct.