THE BEACON PAPERS · No. 17

THE REPAIR GAP

When AI Finds Faster Than Humanity Can Fix — and How Human–AI Co-Repair Can Close the Gap

Nguyen Anh Tuan

Founder and Chief Architect of AIWS

Co-Founder, Co-Chair and CEO, Boston Global Forum

Verona, Italy • 22 October 2026

In the AI Age, security is measured not by how fast we find our flaws, but by how fast we repair them — together.

I. October 21, 2026

On October 21, 2026, the maintainers of PHP, the programming language used by roughly 70 percent of websites whose server-side language is known, released emergency fixes for two critical flaws. Together they became known as PHP2Shell. They could enable unauthenticated remote code execution on affected servers: an outsider, without any login, taking control of the machine. Wiz, one of the research firms that found them, judged that they “rank among the most severe vulnerabilities discovered in PHP, and arguably in software history.”

The flaws had been found by artificial intelligence. Researchers at the security firm Calif, working with OpenAI’s cybersecurity models, reported the first on August 11; Wiz reported the second on August 24. For two months OpenAI, Wiz, Calif and the PHP maintainers worked together in confidence: they built and tested the fix, prepared protections with security and cloud providers, and readied the disclosure. This was responsible work, done well. Then details leaked, the release was moved forward by a day, and within hours, according to Cloudflare’s chief executive, there were millions of attempts to exploit it.

Governments responded as if to a storm. The U.S. National Security Agency advised organizations to take systems offline if they could not update at once. Authorities in Germany, France, Belgium, the Netherlands, Finland and the Czech Republic took government websites offline. Japan’s Prime Minister personally urged administrators to

check their systems. On the same day, attackers planted malicious code in LiteLLM, a widely used open-source gateway to AI models, and it emerged that private keys held by an Indian certificate authority, the keys on which websites rely to prove who they are, had been compromised. The week before, Microsoft had fixed a record 1,411 flaws in a single monthly release, a number it attributes in part to its own use of AI to find them.

Together, these incidents expose a shared structural weakness of the digital world in the AI Age.

II. What AI Has Already Given to Security

It must be said first, and plainly: **artificial intelligence is the most powerful ally that the defense of the digital world has ever had.**

PHP2Shell was found by defenders before it was found by criminals. “To our knowledge,” OpenAI stated, “frontier AI found these vulnerabilities before attackers did.” In July 2025, Google reported that its AI agent Big Sleep had identified a critical flaw in SQLite that attackers were preparing to use, and that it was closed before they could; Google called it the first time an AI agent had directly foiled the exploitation of a vulnerability in the wild. In August 2025, the systems built for DARPA’s AI Cyber Challenge found 54 of 63 planted vulnerabilities in more than 54 million lines of code, produced patches for 43 of them, submitted in an average of 45 minutes within the competition, and discovered 18 real, previously unknown flaws.

Flaws buried in code for twenty years can now be found in weeks. Defenses that once required rare experts can now be extended to every hospital, school and city hall. **The danger is not AI in security. The danger is an imbalance that AI has created, and that AI, guided by capable people and supported by accountable institutions, can help close.**

III. The Repair Gap

The imbalance is this. Discovery now scales with computing power. Repair still scales with people. A flaw is found once, by a machine, in an afternoon. It must be repaired thousands of times: by the volunteer who maintains the code, the vendor who packages it, the ministry, hospital or small business that runs it, each with its own staff, budget and calendar.

This paper names that distance the **Repair Gap**: the time from the earliest documented discovery of a flaw to the moment the people who depend on a system are confirmed to be protected. It has two parts that can be measured. The **Repair Interval** runs from documented discovery to a verified fix. The **Protection Interval** runs from that fix to confirmed protection of the systems people actually use. A temporary mitigation narrows the gap; it does not close it. The gap closes when durable remediation, through a verified fix or the replacement or retirement of the vulnerable component, is implemented and

confirmed. Where a weakness is noticed before it is confirmed as a security vulnerability, both dates are recorded; the time between them belongs to the gap. For most of the history of computing, the gap was hidden, because discovery was slow and exploitation was hard. Neither is true any longer. PHP2Shell shows both changes at once.

First, AI has broken the assumption behind the old rule of triage, under which only flaws proven to be exploitable were treated as urgent. One of the weaknesses behind PHP2Shell had been noticed as early as 2020 and found again in a 2024 audit. It was not fixed then, because no one had shown that it could be exploited. AI agents turned it into a working attack in a matter of weeks. As Wiz put it, the difficulty of exploitation "is no longer a barrier." The absence of a demonstrated exploit can no longer, by itself, justify postponing repair.

Second, secrecy is no longer a reliable shelter. The window in which defenders held the discovery lasted two months, and then it leaked. Amazon's threat sensors recorded an attempt on October 17, four days before the public disclosure. The same capability that lets a defender find a flaw lets an adversary find it again, or learn of it. Every hour between discovery and repair now belongs to whoever moves faster.

There is a third, quieter cost. AI that finds without helping to repair becomes a burden on the very people it means to help. In May 2026, the PHP Foundation reported that its maintainers were receiving a sharply rising volume of vulnerability reports, many "bearing the hallmarks of lazy AI generation," which consumed their time and buried the real issues. A report is not a repair. Discovery delivered to people who cannot absorb it widens the gap it was meant to close.

In the AI Age, the security of a civilization is measured not by how many flaws it can find, but by how fast it can repair them.

IV. The Unseen Commons

Beneath every government website, bank and hospital lies a foundation that almost no one sees: open-source software, written over decades, shared freely, and maintained largely by small groups of volunteers. PHP runs most of the visible web. When the PHP Foundation formed an Ecosystem Security Team in May 2026 to face the rising tide of AI-era reports, its grant provided for one full-time role, for six months, alongside a few further goals that volunteers could not cover.

This is the **Unseen Commons**: the shared foundation of digital life, on which states, corporations and AI laboratories depend, and for whose care almost no one is responsible. The attack on LiteLLM shows that the infrastructure of AI itself now belongs to this commons. The compromise of a certificate authority shows that even the anchors of trust can fail.

When the commons breaks, the cost falls on people who were never in the room: the citizen who cannot reach a public service on the day the ministry goes dark, the small

business owner who has never heard of PHP, the volunteer maintainer asked to carry the security of the world in their spare hours. A civilization cannot rest its daily life on a foundation that it does not see and does not maintain.

V. The AIWS Answer: Human–AI Co-Repair

The answer is not to slow discovery. It is to make repair travel at the speed of discovery. The coordinated work behind PHP2Shell showed that this is possible. What it did not establish is who is responsible when goodwill alone does not organize it, and who carries the repair the last mile, to the ministry, the hospital and the small business where people are actually protected.

AIWS calls its answer **Human–AI Co-Repair**: the method of Human–AI Co-Creation applied to the security of the shared foundations. **Human–AI Co-Repair turns coordinated repair into a continuing institutional responsibility, measured by the protection actually delivered to people.** Its rule is simple. The intelligence that finds a flaw must also help to repair it; human beings who understand the system must judge, authorize and own the repair; and the work is finished not when a patch is published, but when the people who depend on the system are protected.

In practice, every consequential flaw discovered by AI should travel with a **Repair Package**, prepared by AI and verified by people:

- **Evidence —** the flaw, reproduced and confirmed, with its severity and reach.
- **A proposed fix —** a candidate repair, written by AI for those who maintain the code.
- **Proof of the fix —** tests demonstrating that the flaw is remediated and checking for regressions, with remaining uncertainties stated.
- **An explanation —** in plain language, so that the people responsible understand what they are approving.
- **Guidance for operators —** how to deploy the fix, and what to do in the meantime where it cannot yet be applied.
- **Detection —** rules that let defenders recognize attempts to exploit the flaw.

AI drafts; people decide. This division is not a courtesy. Beacon Papers No. 11 established that capability is not authority, and No. 16 that no discovery should enter humanity's knowledge without a path to understanding. A patch that no one understands is the next flaw waiting to be found. In Co-Repair, the machine does the labor at machine speed, and the human being remains the author of the decision. And the package serves the warning, never the reverse: **an urgent, credible warning must not be delayed because a complete Repair Package is not yet available.**

From this follows the principle at the center of this paper: **the power to find carries the duty to help repair.** Responsible disclosure, the ethic of recent decades, asked those who found flaws to report them quietly. The AI Age requires **responsible repair**: whoever holds AI capable of finding vulnerabilities at scale owes the world not only the

report, but a proportionate contribution to verified remediation. A frontier laboratory carries a different measure of that duty than an independent researcher.

VI. Beyond Patching: Five Higher Solutions

Co-Repair closes the gap one flaw at a time. A civilization must also act at the level of its foundations. AIWS proposes five measures.

1. **The Renewal of the Foundations.** PHP2Shell grew out of memory-handling errors deep in code written in C, the oldest and most common class of weakness in the software that runs the world. AI makes it possible, at a scale never before practical, not merely to patch the foundations but to renew them: to read code written decades ago, explain it, and rebuild the most critical components in memory-safe form, verified wherever possible by formal proof, with human engineers understanding and owning every part. Patching repairs yesterday's flaw. Renewal removes whole classes of tomorrow's.

2. **The Repair Gap as a public measure.** What is not measured is not repaired. AIWS proposes that the AIWS Trust Observatory (Beacon Papers No. 13) track the Repair Gap of the software on which public life depends, and of the institutions that run it, reporting both the Repair Interval and the Protection Interval, and distinguishing temporary mitigation from durable remediation. Its watch should extend to the cybersecurity and data security of AI: flaws found by AI, attacks on the infrastructure of AI itself, AI used as an instrument of attack, and the data that AI systems collect and hold. Security is a precondition of trust, and this work is part of AIWS Trust Infrastructure. AIWS Trust Rating should treat the capacity to repair as a mark of trustworthiness. A nation's security can then be seen in the speed with which it heals.

   How the Observatory discloses. When the Observatory finds such a problem, it does not announce it to the public. It works directly with the companies and institutions concerned, proposes ideas for a solution, and presses for the flaw to be remedied and the matter resolved; when that is complete, it publishes. Because secrecy is fragile, this confidence exists to shorten the time to repair, never to lengthen the time a flaw stays hidden. Deadlines are set in proportion to the risk, after the responsible party has had the opportunity to respond. If a party fails to act within that time, the Observatory makes public that the problem exists and who has not acted, disclosing nothing beyond what users need to protect themselves. Where users must act at once and the responsible party has not warned them, the Observatory issues an actionable warning that withholds exploit-enabling detail, as Section V requires.

3. **The Repair Dividend.** Those who draw the greatest value from the Unseen Commons — frontier AI laboratories, cloud providers, large enterprises and governments — should return a share of their AI capacity and expertise to its care. A

defender's window should not be a favor granted by whoever happens to find a flaw first. It should be a standing commitment, in which defensive AI serves the shared foundations before it serves anything else.

4. **Civilian Immunity of the Commons.** A flaw in software on which civilian life depends is a danger to everyone who uses it. AIWS proposes that states and companies commit that vulnerabilities in these shared foundations, however discovered, be delivered for repair and never stockpiled for attack. A hospital in one country runs the same code as a ministry in another. To hold such a flaw in reserve is to hold every civilian population hostage at once.

5. **A generation of defenders.** No measure endures without the people to carry it. The repair of the digital world must become a human vocation of the AI Age, taught in universities and practiced in every institution. AIWS proposes that AIWS University and the Human Advancement Test (Beacon Papers No. 15) help form defenders who become more capable with AI, not dependent upon it: people who can read what the machine has found, judge what it proposes, and stand behind the repair.

VII. Proposed Articles for the Constitution for Humanity

To give these measures a lasting foundation, AIWS proposes the following two articles for the AIWS Constitution for Humanity in the AI Age:

PROPOSED ARTICLE I

The Security of Humanity's Shared Foundations

1. The software and systems on which civilian life depends are a common inheritance of humanity.

2. Whoever discovers a vulnerability in them, by human or artificial means, shall convey it securely to those responsible for its repair, and shall not retain it for attack.

3. Those who hold the capacity to discover vulnerabilities at scale bear a duty, proportionate to that capacity, to contribute to their verified repair.

4. Repairs of these foundations shall remain under accountable human authority; automated repair is permitted only when explicitly authorized, with verification and recovery safeguards proportionate to risk.

PROPOSED ARTICLE II

The Security of Persons and Their Data

1. Every person has the right to the security of their personal data and of the digital systems on which their life depends.

2. Whoever collects or holds personal data, or operates AI systems that use it, bears a duty to protect it, to inform those affected when it is compromised, and to repair the breach without delay.

3. No state, organization or person shall attack the infrastructure on which civilian life depends, by artificial intelligence or by any other means.

4. AI systems shall be protected against tampering, theft and poisoning in proportion to the authority entrusted to them.

5. Security shall never be invoked to justify mass surveillance or cognitive control. A security that places every citizen under watch is not the protection of the people; it is power over them.

VIII. A Civilization That Can Repair Itself

Civilizations have never endured because nothing in them broke. They endured because they learned to repair: their roads and bridges, their laws, their institutions, and their trust in one another. The AI Age adds a new demand to this old wisdom. Artificial intelligence will find more of our flaws, and find them faster, than any generation before us could have imagined. Whether that becomes a danger or a gift depends on one question: will we build a civilization able to repair itself at the speed of its own discovery?

AIWS Civilization answers with Human–AI Co-Creation. Let AI find. Let AI help repair. Let human beings understand, decide and renew. Then every flaw that AI uncovers becomes not a threat to the foundations of our common life, but a step in making them stronger than we found them.

BEACON PRINCIPLE No. 17

AI may find.
AI must help repair what it finds.
Humanity must understand, decide, and renew.

FIND TOGETHER. REPAIR TOGETHER. RENEW TOGETHER.

SOURCES

Beacon Papers

No. 11, No Power Without a Mandate (Sept 2026). https://bostonglobalforum.org/news/beacon-papers-no-11-no-power-without-a-mandate/

No. 13, AI Power Must Not Operate Unseen (Sept 2026).

No. 15, The Human Advancement Test (Oct 13, 2026). https://bostonglobalforum.org/wp-content/uploads/Beacon_Papers_No15_The_Human_Advancement_Test_EN.pdf

No. 16, No Discovery Without Understanding (Oct 18, 2026).

Other sources

Help Net Security, "PHP2Shell: Critical PHP vulnerabilities patched, update now!" (Oct 21, 2026). https://www.helpnetsecurity.com/2026/10/21/php2shell-php-vulnerabilities-patched-cve-2026-109618-cve-2026-109619/

SecurityWeek, "PHP2Shell Patches Land as First Exploitation Attempts Hit Sensors" (Oct 2026). https://www.securityweek.com/php2shell-patches-land-as-first-exploitation-attempts-hit-sensors/

CyberScoop, "PHP releases emergency fixes for 2 critical vulnerabilities" (Oct 2026). https://cyberscoop.com/php-emergency-patches-php2shell-vulnerabilities/

Al Jazeera, "World leaders warn of security flaws found in PHP programming language" (Oct 22, 2026). https://www.aljazeera.com/news/2026/10/22/world-leaders-warn-against-security-flaw-found-in-php-computer-programming

PBS NewsHour / AP, "AI discovered a software vulnerability that could impact 70% of websites worldwide" (Oct 2026). https://www.pbs.org/newshour/nation/ai-discovered-a-software-vulnerability-that-could-impact-70-of-websites-worldwide-the-rush-is-on-to-patch-it

Wiz, "Breaking Down PHP2Shell with Sagi Tzadik & Thai Duong." https://www.wiz.io/crying-out-cloud/breaking-down-php2shell-with-sagi-tzadik-and-thai-duong

TechCrunch, "The internet is having a very bad cybersecurity day" (Oct 21, 2026). https://techcrunch.com/2026/10/21/the-internet-is-having-a-very-bad-cybersecurity-day/

BleepingComputer, "Microsoft October 2026 Patch Tuesday fixes 1,411 flaws, 3 zero-days" (Oct 13, 2026). https://www.bleepingcomputer.com/news/microsoft/microsoft-october-2026-patch-tuesday-fixes-1-411-flaws-3-zero-days/

The PHP Foundation, "Announcing the Ecosystem Security Team at The PHP Foundation" (May 18, 2026). https://thephp.foundation/blog/2026/05/18/announcing-ecosystem-security-team/

DARPA, "AI Cyber Challenge marks pivotal inflection point for cyber defense" (Aug 8, 2025). https://darpa.mil/news/2025/aixcc-results

The Hacker News, "Google AI 'Big Sleep' Stops Exploitation of Critical SQLite Vulnerability Before Hackers Act" (July 16, 2025). https://thehackernews.com/2025/07/google-ai-big-sleep-stops-exploitation.html

TechTarget, "PHP2Shell disclosure unravels after leaks forced early patch drop" (Oct 2026). https://www.techtarget.com/cybersecurity/news/366652613/PHP2Shell-disclosure-unravels-after-leaks-forced-early-patch-drop

SiliconANGLE, "PHP2Shell attacks began four days before patches, Amazon says, as attempts hit millions" (Oct 22, 2026). https://siliconangle.com/2026/10/22/php2shell-attacks-began-four-days-before-patches-amazon-says-as-attempts-hit-millions/

Wiz, "PHP2Shell: AI uncovers critical RCE vulnerabilities affecting 70% of the Internet" (Oct 21, 2026). https://wiz.io/blog/php2shell-php-rce

PPC Land, "OpenAI says GPT-5.6 Sol found a PHP flaw before attackers did" (Oct 2026). https://ppc.land/openai-says-gpt-5-6-sol-found-a-php-flaw-before-attackers-did/

Insurance Business, "NSA warns of critical website flaw as Wikipedia, European government websites taken offline" (Oct 2026). https://www.insurancebusinessmag.com/us/news/cyber/nsa-warns-of-critical-website-flaw-as-wikipedia-european-government-websites-taken-offline-594432.aspx

Reuters, "FACTBOX-Websites taken offline over PHP risks," via Devdiscourse (Oct 2026). https://www.devdiscourse.com/article/international/3995980-factbox-websites-taken-offline-over-php-risks

The Next Web, "European governments pull websites offline over two critical PHP flaws" (Oct 2026). https://thenextweb.com/news/europe-governments-websites-offline-php-security-flaws

Tech-ish, “Japan’s PM, the NSA and the FBI tell websites to update PHP or go offline” (Oct 21, 2026). https://tech-ish.com/2026/10/21/japans-pm-the-nsa-and-the-fbi-tell-websites-to-update-php-or-go-offline/

Datadog Security Labs, analysis of the compromised LiteLLM releases on PyPI and Docker (Oct 21, 2026). https://securitylabs.datadoghq.com/articles/litellm-compromised-pypi-docker/

CSO Online, “Google, Mozilla revoke eMudhra intermediates after private key compromise” (Oct 2026). https://www.csoonline.com/article/4238338/google-mozilla-revoke-emudhra-intermediates-after-private-key-compromise.html

Verona, Italy • 22 October 2026