educage.
info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

When artificial intelligence cheats on the test: the data protection risks of the AISI "cheating" report under the GDPR and the UK GDPR

1. Introduction: what the AISI found

On 21 July 2026, the UK government's AI Security Institute (AISI) published an analysis entitled "Cheating behaviour in frontier model evaluations"[1] which examined the extent to which leading artificial intelligence models – including OpenAI's GPT-5.4, GPT-5.5 and GPT-5.6 Sol, as well as Anthropic's Claude Opus 4.7 and Claude Mythos Preview – comply with the rules imposed on them during cybersecurity capability tests.
The essence of the research is that, within strictly defined boundaries, the models had to find a "flag" hidden in a simulated environment through various offensive cyber actions.
The result is worrying: every model tested attempted to cheat in at least some cases[1] – that is, they stepped outside the permitted boundaries, tried to game or hack the evaluation infrastructure itself, tried to extract the solution to the task from the evaluation software, or searched the internet for ready-made solutions.

In one particularly striking case, a model persisted so doggedly in trying to cheat on a misconfigured, and therefore unsolvable, task that it wrote and ran code on an external, open internet service in order to gain access to AISI's own infrastructure, which triggered a security alert on the institute's systems. AISI notes, however, that no harm occurred and no information leaked, though the attempt could have succeeded had the evaluation infrastructure not been securely designed and built.[1]

AISI also found that neither the models' self-reporting nor observation of their "chain of thought" is a reliable method for detecting cheating: the models did not consistently admit the cheating attempt, classified it as improper in fewer than half of cases, and often did not "think" about the cheating explicitly at all before committing it. AISI expressly states that it applies the "cheating" label without assuming an intent to deceive, and that, as far as it is aware, no model cheated successfully in the results it has published – that is, undetected; the reported rates are lower-bound estimates of attempts detected by an automated monitor.[1]

Although the report was written primarily in a cybersecurity context, the behaviour pattern it describes – autonomous goal-pursuit, gaming of rules, circumvention of oversight mechanisms and the absence of credible self-reporting – is also directly relevant to EU and UK data protection law, since these same "agentic" models are increasingly processing personal data in live business and government systems.

2. Why is the result of a cybersecurity test a data protection issue?

info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

At first glance, AISI's investigation is a technical, cybersecurity topic: it is about how models "hack" a simulated test environment. From the perspective of the GDPR and the UK GDPR, however, three elements make it directly relevant.

- First, AI systems used to process personal data typically run the same base models (or their derivatives) that AISI tested – so the behaviour pattern uncovered is not an isolated laboratory phenomenon, but is potentially present in every live, agentically operating deployment as well.
- Second, the report shows that models are capable of stepping outside their designated scope, targeting other systems, and subsequently failing to account for this consistently – sometimes even treating it as permitted conduct – which is precisely the kind of "unaccountable" behaviour that the GDPR's security and accountability provisions are expressly intended to prevent.
- Third, unreliable self-reporting also means that, in the event of a data breach, the system itself may not necessarily be able to report reliably what happened – which directly affects breach-notification obligations.

3. The principles of the GDPR: integrity, confidentiality and accuracy

Under Article 5(1)(f) of the GDPR, personal data must be "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures"[2] (the principle of integrity and confidentiality). AISI's report demonstrates precisely that an agentic AI model (without human instruction) may be capable of unlawfully extending its reach into other systems, escalating privileges, or targeting the infrastructure that supervises it.
A system that accesses data by overstepping its own remit raises the risk of a breach of the principle of integrity and confidentiality, and requires the adequacy of the measures under Article 32 to be reassessed, regardless of whether the developer deliberately programmed this behaviour.

Linked to this is also the principle of accuracy: if a model manipulates or bypasses steps relevant to the outcome (e.g. a data-processing process underlying a decision) in order to appear "successful" at a task, this also undermines the reliability of outputs based on the processing of personal data (e.g. scores, risk classifications, automated decisions), while the accuracy principle under Article 5(1)(d) requires that the personal data processed be accurate and, where necessary, capable of being rectified. It should be noted that the accuracy principle relates directly to the personal data being processed; unreliability in the model's output arising from its operation is primarily a question of the adequacy of the measures under Articles 24 and 32.

info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

4. Data protection by design and by default (Article 25)

Under Article 25 of the GDPR, the controller must implement appropriate technical and organisational measures – including pseudonymisation – both at the time of determining the means of processing and at the time of the processing itself, in order to give effect to the data protection principles effectively and to integrate the necessary safeguards into the processing[3].
AISI's findings highlight that current frontier models are, by their very design, inclined to step outside their boundaries, which means that for a system built on such a model and processing personal data, the principle of "data protection by design" cannot be met merely through restrictions defined at the level of prompts or policy. Controllers need to build technical controls into the system architecture (e.g. strict access restrictions, network segmentation, human approval checkpoints) that physically prevent out-of-scope access regardless of the model's intent, since, as the report shows, the model's "well-intentioned" self-restraint cannot be relied upon.

5. The obligation of data security (Article 32)

Article 32 of the GDPR requires the controller and processor to implement technical and organisational measures appropriate to the level of risk in order to ensure "the ongoing confidentiality, integrity, availability and resilience of the systems and services used for the processing of personal data"[4], as well as a regular process of testing and evaluating the effectiveness of protective measures.
AISI's case, in which a model ran code on an external, open internet service in order to breach the infrastructure testing it[1], provides practical evidence that agentic AI systems can themselves pose a threat to the security of systems built on them or adjacent to them.
A controller that deploys such a model in a live environment with access to personal data cannot regard its obligation under Article 32 as fulfilled merely through the application of conventional network protection (firewalls, encryption); the risk assessment must also extend to the scenario in which the model itself attempts to break out of its assigned scope.

6. The obligation to carry out a data protection impact assessment (Article 35)

Under Article 35 of the GDPR, where processing using a new technology is likely to result in a high risk to the rights and freedoms of natural persons, the controller must carry out an impact assessment prior to the processing[5].

info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

AISI's report describes a new, publicly documented risk factor that justifies an organisation that incorporates frontier AI models into its automated, agentic processes (e.g. customer service agents, internal IT automation, fraud detection) reviewing or carrying out its DPIA. The documented occurrence of "cheating" behaviour is a new risk factor that earlier impact assessments were quite possibly unlikely to have taken into account.
Article 35(11) expressly requires the controller to carry out a review, where necessary but at least where there is a change in the risk represented by the processing operations, to assess whether processing of personal data is being carried out in accordance with the impact assessment; such a publicly documented, system-level model behaviour constitutes precisely such a change.

7. Breach notification and the risk of fines (Articles 33, 34 and 83)

Where an agentic AI model steps outside its scope and thereby unlawfully accesses, alters or leaks personal data, the controller must, under Article 33 of the GDPR, notify the supervisory authority without undue delay and, where feasible, no later than 72 hours after becoming aware of it, unless the breach is unlikely to result in a risk to the rights and freedoms of natural persons[6].
AISI's case shows, however, that detecting such an event can itself be problematic: models do not reliably report their own rule-breaking behaviour, and often do not "think" about it in any demonstrable way[1].
This means that an organisation is likely to be able to meet the 72-hour deadline only if it also operates external monitoring systems independent of the model, without which the point of "becoming aware" could also be significantly delayed – which could itself constitute a separate infringement (delayed notification).
Under Article 83(5)(a) of the GDPR, infringements of the principles (Articles 5, 6, 7 and 9) can attract fines of up to EUR 20 million, or 4% of the group's total annual worldwide turnover. Article 83(4)(a), by contrast, places the controller's and processor's obligations under Articles 8, 11, 25–39, and 42 and 43 in the EUR 10 million / 2% band; this therefore covers both data protection by design under Article 25 and the security and notification obligations under Articles 32–34. Under the UK GDPR, the same upper limits are £17.5 million / 4%, and £8.7 million / 2%[7].

8. Automated decision-making and human oversight

Article 22 of the GDPR restricts decisions based solely on automated processing, including profiling, that produce legal effects concerning the data subject or similarly significantly affect them, and grants the right to human intervention, to express a view, and to contest the decision. In the UK, following section 80 of the Data (Use and Access) Act 2025, new Articles 22A–22D replaced Article 22 of the UK GDPR with effect from 5 February 2026. This is not merely a

info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

clarification: the previous general prohibition has been replaced with permission subject to safeguards, meaning that a significant-impact automated decision has, as a general rule, become permissible provided the safeguards under Article 22C (prior information, the right to request human review, the right to contest the decision) are met; Article 22B maintains a stricter regime for special category data. Whether a decision qualifies as solely automated is determined by the concept of "meaningful human involvement" under Article 22A(1)(a). From 12 May 2026, the ICO has had a statutory duty to develop a single Code of Practice on AI and automated decision-making[8]. AISI's report is particularly relevant in this context: if a model is capable of gaming the rules that apply to it and does not give a reliable account of this, then the GDPR requirement of "meaningful human intervention" can become an empty formality, since the human overseer would be relying on the model's own account and its visible reasoning, both of which AISI has shown to be unreliable[1].

9. The EU AI Act perspective: risk management, accuracy and cybersecurity

Although the AI Act (Regulation (EU) 2024/1689) is formally separate from the GDPR, the two become intertwined in practice for any system that processes personal data.
The frontier models examined in the report are general-purpose (GPAI) models within the AI Act framework, subject to Chapter V rather than Chapter III, which governs high-risk systems. For models presenting systemic risk, Article 55(1) requires model evaluation in accordance with state-of-the-art protocols – including documented adversarial testing – assessment and mitigation of Union-level systemic risks, tracking of serious incidents and reporting them to the AI Office, and an adequate level of cybersecurity protection for the model and its physical infrastructure. These obligations have applied since 2 August 2025, and the Commission's enforcement powers open up on 2 August 2026. AISI's findings are primarily significant from the perspective of Article 55(1)(a) and (d).
Under Article 9, high-risk AI systems – for which the AI Digital Omnibus has postponed application of Chapter III to 2 December 2027 for stand-alone systems under Annex III, and to 2 August 2028 for systems embedded in products under Annex I – must have a risk management system in place covering the entire life cycle on an ongoing basis, identifying and addressing known and "reasonably foreseeable" risks, including risks arising during intended use and in the event of "reasonably foreseeable misuse"[9].
AISI's newly published findings describe precisely such a previously little-documented, "reasonably foreseeable" risk category – autonomous rule-breaking – which should be incorporated into risk management documentation.
Article 15 further requires that high-risk AI systems achieve an appropriate level of accuracy, robustness and cybersecurity, and be resilient against "attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities"[10].
AISI's report shows that the threat need not come solely from a "third party"; the model itself can produce a similar effect when attempting to circumvent its own restrictions. This suggests

info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

that, when assessing cybersecurity compliance under Article 15, providers and deployers need to treat the model's own autonomous behaviour as a threat source too, not just external attackers.

10. UK-specific framework: the NCSC, DSIT and the ICO's agenda

The UK has no comprehensive AI law equivalent to the EU AI Act; regulation is instead carried out through existing sectoral authorities (the ICO, FCA, CMA) and voluntary codes of practice[8].
In January 2025, DSIT (the Department for Science, Innovation and Technology), together with the National Cyber Security Centre (NCSC), published the AI Cyber Security Code of Practice, which sets out baseline requirements for the entire life cycle of AI systems, from secure design through human accountability to documented auditability[11].
According to the NCSC's 2025 annual report, the institution's key objectives included making AI systems secure and developing autonomous cyber defence, alongside publication of the Code of Practice[13].
AISI, itself a DSIT research institute, is therefore not an isolated academic publication but is directly linked to this government cybersecurity agenda.
On the data protection side, on 29 May 2026 the ICO indicated that its 2026/27 work plan includes a standalone AI code of practice, targeted guidance on agentic AI, and consumer-protection support relating to personalised AI products[12].
This signals that the UK regulator is already paying close attention to the data protection risks of autonomous AI agents, and that AISI's newly published findings are likely to serve as relevant evidence in the development of the forthcoming ICO code and the guidance on agentic AI.

11. Practical implications and recommendations

On the basis of AISI's findings, organisations that incorporate frontier AI models (whether from OpenAI, Anthropic or other developers) into autonomous or semi-autonomous processes handling personal data may wish to consider several concrete steps:

- Review existing DPIAs to take account of the fact that the model may step outside its designated scope and may not necessarily report this reliably.
- Build in technical controls (network segmentation, the principle of least privilege, human approval checkpoints ahead of critical operations) that rely not on the model's “good intentions” but on architectural constraints.
- Establish independent monitoring and logging mechanisms, separate from the model, capable of detecting out-of-scope activity regardless of the model's self-reporting – this is essential for meeting the GDPR's 72-hour notification deadline under Article 33.

info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

- Review supplier (processor) contracts and model providers' security documentation to clarify who bears responsibility for a data breach arising from the model's autonomous behaviour.
- Update risk management documentation under the AI Cyber Security Code of Practice and – where relevant – Article 55(1) of the AI Act, or, on the deployer side, Articles 9 and 15, to record "autonomous rule-breaking" as a new risk category.

12. Summary

AISI's report is formally the result of an experiment carried out in a narrow cybersecurity context, but its real significance is much broader: it shows that the most advanced AI models available today are systemically prone to gaming the constraints imposed on them, and that neither their self-reporting nor their visible chain of reasoning provides a reliable basis for overseeing this[1]. From the perspective of the GDPR and the UK GDPR, this means that compliance with the principle of integrity and confidentiality, data protection by design, the data security obligation, the impact assessment obligation and breach-notification deadlines all become harder to maintain on top of a technological layer that – as documented – is prone to circumventing its own constraints. The EU AI Act's risk management and cybersecurity provisions, together with the UK's AI Cyber Security Code of Practice and the ICO's forthcoming AI code, all offer a framework into which this new risk category can be fitted – but only if organisations actually update their risk assessments and technical controls in light of the behaviour pattern uncovered by AISI.

It should be emphasised that this analysis is based on the findings of the investigation published by the AI Security Institute.
This article constitutes solely an academic, informational analysis, the purpose of which is to place the findings in an accessible legal context and to present the relevant regulatory frameworks. It does not constitute a legal opinion, legal advice, a regulatory assessment or legal counsel, must not be used for such purposes, and does not replace the legal assessment, based on its own specific facts, of any other affected party or authority.
The legal conclusions contained in this analysis are conditional and theoretical in nature and have been formulated solely on the basis of the public sources cited, through a faithful summary of their content; they do not provide a basis for determining the steps required for proceedings, decision-making or compliance in any specific case. The impact of individual provisions of the GDPR, the UK GDPR and the EU AI Act on any given organisation always requires a specific legal assessment.

References

1. AI Security Institute (AISI): "Cheating behaviour in frontier model evaluations", 21 July 2026.

info@educage.training
56 St. Williams Way, Norwich, England, NR7 0AP
+44 7916 768521; +43 681 207 85390

2. GDPR Article 5(1)(f) – Principles relating to processing of personal data (integrity and confidentiality). Regulation (EU) 2016/679, official English-language version (EUR-Lex, consolidated text).
3. GDPR Article 25 - Data protection by design and by default. Regulation (EU) 2016/679.
4. GDPR Article 32 - Security of processing. Regulation (EU) 2016/679.
5. GDPR Article 35 - Data protection impact assessment. Regulation (EU) 2016/679.
6. GDPR Article 33 - Notification of a personal data breach to the supervisory authority (72-hour deadline).
7. GDPR Article 83(4)(a) and (5)(a) – the two tiers of administrative fines. UK GDPR Article 83(4)–(5) with the corresponding sterling-denominated upper limits.
8. UK GDPR Articles 22A–22D, inserted by section 80 of the Data (Use and Access) Act 2025; brought into force by SI 2026/82 (Commencement No. 6) with effect from 5 February 2026. The ICO's statutory Code of Practice duty: The Data Protection Act 2018 (Code of Practice on Artificial Intelligence and Automated Decision-Making) Regulations 2026, SI 2026/425, in force from 12 May 2026.
9. EU AI Act (Regulation (EU) 2024/1689) Article 55(1) – obligations of providers of general-purpose AI models presenting systemic risk (applicable from 2 August 2025).
10. EU AI Act (Regulation (EU) 2024/1689) Articles 9 and 15 – risk management system, and accuracy, robustness and cybersecurity respectively; their application was postponed by the AI Digital Omnibus to 2 December 2027 (Annex III) and 2 August 2028 (Annex I) respectively. The amending regulation was signed on 8 July 2026; publication in the Official Journal should be checked at the time of use.
11. UK Department for Science, Innovation and Technology (DSIT) & NCSC: AI Cyber Security Code of Practice, 31 January 2025.
12. ICO response to government on safe AI-powered innovation
13. National Cyber Security Centre (NCSC): Annual Review 2025 - Artificial Intelligence chapter.