Comments by Milestone Systems on Guidelines 02/2026 on Anonymisation

We welcome the publication of the draft Guidelines 02/2026 and appreciate the EDPB's commitment to providing a consistent and up-to-date approach to anonymisation.

Milestone Systems develops video management software. Since acquiring the Berlin-based anonymisation company brighter AI in 2025, we also provide technology that detects faces and licence plates in images and video and either blurs them or replaces them with synthetic alternatives. Our customers use it for footage from vehicles, robots and public spaces, including data used for AI development.

We particularly welcome the relative, entity-specific approach to identifiability (paragraphs 6–12), the clarification that the likelihood of identification must be insignificant rather than zero (paragraph 22), and the recognition that anonymisation close to capture, with prompt deletion of the originals, can limit the impact on the individuals recorded (paragraph 39). Anonymisation is itself processing that requires a legal basis (paragraph 38) and cannot make unlawful collection lawful. However, the framework is largely built around tabular data. Video raises distinct questions, on which we would like to submit the following observations.

Our observations are grouped into three areas that follow the steps of the anonymity assessment itself, namely **identifiability in video**, which concerns what makes an individual identifiable in continuous footage, **evidence and method**, which concerns how effective anonymisation can be demonstrated, and **risk threshold and recipients**, which concerns where the line of acceptable risk lies and for whom it should be assessed.

To support a consistent application of the Guidelines to video, we outline in the Annex a provisional methodology for testing anonymisation in visual data. It is offered as one way to produce the documentation called for in paragraph 41, not as a proposed legal threshold.

Group 1: Identifiability in video

1) Clarify when temporally consistent video amounts to singling out.

Many temporal perception tasks require each person to appear coherently across frames. Detailed tracks, and the wider record they form part of (Glossary, "Record"), will therefore often contain unique combinations of attributes, so that the No Record Isolation criterion will frequently fail (paragraphs 55 and 56). Under paragraphs 98 to 100, the decisive question then becomes whether an isolated record can be matched to the corresponding attributes of the person. Example 24 walks through this further analysis for a survey, but the draft contains no equivalent for video.

We therefore suggest adding a worked example that contrasts offline training footage, in which tracks are unique but cannot be matched to anyone with means reasonably likely to be used by the relevant entities, with live or near-live footage of a known place and time, in which a track allows

someone to act towards the person. It would also help to confirm that consistency within one recording does not by itself prevent anonymity, while consistency across recordings, or combinations of clothing, gait, vehicle, time and location, can enable singling out or linkage and must therefore be assessed (paragraphs 24 and 25, Example 5). We further suggest that the final Guidelines indicate how features remaining after faces and licence plates have been removed, such as gait, build, clothing, scene density and accompanying persons, should be weighed in that assessment, in particular depending on whether a matching source is reasonably available.

2) Provide clarity on retained attributes, annotations and downstream models.

Anonymised training video must retain some information to remain useful, such as pose, motion, gaze direction, facial expression or approximate age. Annotations, timestamps and location tags often accompany the footage, and models trained on it may later be shared. Paragraph 67 limits the No Inference criterion to inferences that are specific and meaningful, paragraphs 70 and 71 explain how to show whether an inference relies on the data of an individual, and paragraph 82 notes that synthetic data and AI models can enable such inferences.

A worked example could clarify how task-relevant attributes such as pose, motion, gaze and expression may be retained consistently with the No Inference criterion, provided that the released data, together with other information reasonably likely to be available to the relevant entities, does not permit a specific and meaningful inference about an identified or identifiable individual. We further suggest confirming that annotations and metadata accompanying footage are part of the given data, assessed together with the context in which the data is produced and released (paragraphs 29(b) and 68), and clarifying when an evaluation of downstream models is necessary, what it should cover and who is responsible for it, consistent with Opinion 28/2024. In our view, where training footage has been shown to be anonymous before training, a separate anonymity assessment of the resulting model should generally not be required, and model evaluation under paragraph 82 and Opinion 28/2024 should mainly concern models trained on personal data.

Group 2: Evidence and method

3) Recognise structured empirical testing as evidence of effective anonymisation.

For video, the difficulty of the assessment lies in showing that identifying information is reliably protected across complete appearances and across everything released with them. Faces and licence plates that are missed or insufficiently transformed are a major source of failure, and identifying information can also remain elsewhere in the data package. Measuring per frame understates this risk, since a tool that missed 1 face in 100 frames, independently and at random, would leave about 78% of five-second appearances (150 frames at 30 frames per second) exposed at least once. The draft requires anonymisation to be effective for every included individual (paragraph 36) and testing to be documented (paragraph 41), and judges re-identification techniques by whether they produce sufficiently accurate and confident results (paragraph 87). Paragraph 84 encourages scientifically grounded expertise, but the draft does not explain how sampled testing can support a conclusion about a large dataset.

We therefore suggest recognising that independent, representative testing with stated uncertainty, measured per appearance rather than per frame, can support an anonymity assessment, and setting out expectations for sampling difficult conditions, reporting confidence bounds, checking subgroups and outliers, correcting or removing detected failures before data is used or released, and reassessing when tools, settings or context change (paragraphs 35 and 42). We further suggest clarifying how controllers and independent recipients may rely on independently validated upstream evidence within its stated scope, supplementing it with the assessment that their own data, access arrangements and use require. *We would be glad to contribute empirical measurement work on this question, as paragraph 84 invites.*

4) Confirm that the criteria apply equally to all anonymisation methods.

Recognition models can identify faces in some blurred and pixelated images (McPherson et al., 2016). A visibly blurred face is therefore no evidence of protection, just as a realistic synthetic face is no evidence of exposure. Some operational standards nevertheless test how well a face is obscured instead of whether the person can still be identified. China's recommended standard GB/T 44464—2024, for example, checks whether an anonymised face is still detected as a face (Annex B.6.1.1) and whether eye, nose and mouth contours can be made out (B.6.2.2). Tests of this kind can reject a realistic replacement even where the source identity is well protected, and on their own they do not establish whether anyone remains identifiable.

The final Guidelines should confirm that the criteria are outcome-based, so that the same standard of protection applies to masking, blurring, pixelation and synthetic replacement, with tests suited to the failure modes of each method.

Group 3: Risk threshold and recipients

5) Clarify how residual failures and attack success should be interpreted.

Even a protection rate of 99.9% per appearance leaves roughly 1,000 exposed appearances in every million. Such a rate describes the process, and on its own it cannot show that every individual in a dataset is protected, as paragraph 36 requires. At the same time, a recognition rate slightly above chance, where an attacker cannot tell which matches are correct, may not amount to identification under paragraph 87 and Example 20, while an average result can conceal a small group of individuals who remain fully recognisable. Moreover, sampled checks can find and remediate failures in the sample, but in a large dataset some failures will statistically remain undetected in the unsampled part.

The final Guidelines could clarify how attack success, confidence and attribution should be interpreted under paragraph 87 and Example 20, without aggregate results overriding the reliable identification of particular individuals. We further suggest clarifying how paragraphs 22 and 36 apply to such undetected residual failures. In our view, their likelihood of identification can be insignificant where the failure rate is demonstrably below a narrow bound, systematic weaknesses in particular subgroups or conditions have been excluded, detected failures have been remediated, and the access context, such as the absence of raw exports, coarsened time and location data and

a controlled environment, prevents an isolated undetected failure from being searched for or attributed to a known person.

6) Demonstrate the assessment along a multi-party video pipeline.

Video used for AI development often passes through several hands: the organisation that records it, an anonymisation service acting on its instructions, a library operator that curates and annotates it, and developers who train models on it. The draft explains the relevant perspectives through two separate examples (Examples 3 and 4), but not how they combine along such a chain. We agree that an anonymisation service acting on the instructions of a controller is assessed from the perspective of that controller (paragraph 15). Two questions nevertheless remain open further down the chain: when a library operator acts for its own purposes, and when footage can be anonymous for an independent developer while remaining personal for the organisation that recorded it (paragraphs 7, 12 and 37).

Paragraph 21 already refers to a reasonably likely ability to transfer the data and to identification through means reasonably likely to be used. Its further condition, that identification "cannot be ruled out", could still be read as requiring a controller to exclude every conceivable access chain, which is difficult to reconcile with the insignificant-likelihood standard in paragraph 22. The EDPB's October 2026 summary also words the first condition differently, referring to an entity "reasonably likely to transfer" the data, whereas paragraph 21 refers to one able to do so.

We therefore suggest adding a worked example with a footage owner, an anonymisation service acting as processor, a library operator and an independent developer, showing from whose perspective each step is assessed and when anonymity for the developer can be concluded. The example could show two variants, one in which the recording organisation retains the original footage, so that the data remains personal for it while being anonymous for the independent developer, and one in which it anonymises close to capture and promptly deletes the originals and any linkable metadata, so that the data can become anonymous for the recording organisation itself (paragraphs 12 and 39). We further suggest clarifying how the "cannot be ruled out" wording applies consistently with paragraphs 9 and 22, distinguishing the ability to transfer data from the likelihood of a particular chain of access and identification, and explaining what evidence suffices to conclude that this likelihood is insignificant.

7) Clarify how access controls narrow the relevant entities.

Controlled development platforms restrict who can access video and what can be exported. Paragraph 30 lists many potential entities, and footnote 25 notes that not all of them need to be considered systematically in each case. The draft accepts that legal prohibitions generally hold unless there is concrete evidence otherwise (paragraphs 32 and 33), treats contracts as a complement to technical measures (paragraph 34), limits access to authorised users where appropriate security controls apply (paragraph 89), and adds that such measures are not sufficient by themselves (paragraph 90).

To help organisations concentrate on credible access chains, we recommend adding a worked example for a controlled-access video library, showing which entities remain relevant once access is restricted, logged and limited in what can be exported, what evidence shows that technical,

organisational and contractual controls are effective, and how the safety margins recommended in paragraph 88 should be set.

Conclusion

Video is increasingly used to develop and train AI systems, and effective anonymisation is essential to doing so in a way that protects the individuals recorded. We respectfully invite the EDPB to consider these clarifications in the preparation of the final version of the Guidelines. The worked examples requested above could also be combined into one video example that follows footage through collection, transformation, library enrichment and developer access, with variants for the availability of originals, export permissions, metadata and public disclosure. We remain available to contribute video examples, test data and technical expertise on the anonymisation of video, as paragraph 84 invites.

Annex: Proposed methodology for testing anonymisation in visual data

This annex outlines a voluntary, independent testing programme that we propose to pilot with users, laboratories and other tool makers. It produces evidence about tools and deployments, but does not decide whether a dataset is anonymous. All figures are illustrative, and the methods below remain open questions for the pilot. Here, users are organisations that run anonymisation tools, such as vehicle makers and cities, and tool makers are the companies that build them. As with Euro NCAP crash tests, makers know the rules, but the programme chooses what is tested.

A1) Keep tool performance, deployment context and the anonymity conclusion apart.

Tool performance, established by a neutral laboratory, concerns how reliably a given tool version and configuration protected faces and licence plates in the tested conditions. The deployment and disclosure context, documented by the user and spot-checked by an auditor, concerns which data package reaches whom, under what access controls, and with what additional information reasonably likely to be available. The anonymity conclusion, documented by the controller under paragraph 41, concerns whether that data package is anonymous for the relevant entities in that context.

Internal development, controlled sharing and public release are therefore disclosure scenarios, each with its own assessment, and none of them is anonymous or personal by default. A tool result can be reused across customers within its tested scope, while a dataset assessment travels with the dataset and each recipient adds what is specific to its own access and use.

A2) Apply one standard of protection, with tests suited to each method.

Every method has to show that the relevant faces, licence plates and other identifiers were found (coverage), that enough identifying information was removed (transformation), that protection holds across the whole appearance, that other released information does not undo it, and that the evidence matches the tool version and settings actually used (auditability).

Opaque masks typically fail where a mask is too small or misplaced, or where identifiers such as hair, tattoos or clothing lie outside it, which calls for coverage and size checks and a review of the whole scene. Blurring and pixelation fail where they are too weak, leave residual detail usable by recognition models or are inconsistent across frames, which calls for machine and human recognition attacks on the output (A4) and strength checks by face size. Synthetic replacement fails where the source identity leaks into the replacement, where the replacement resembles another real person, where a persistent synthetic identity links appearances, or where a face is missed, which calls for recognition attacks against the source (A4), the scoped tests in A5 and the coverage evidence in A6.

A3) Measure protection per appearance rather than per frame.

The unit of measurement is the appearance: one continuous presence of one person or vehicle in one recording. An appearance counts as exposed if one or more frames leave the person or vehicle identifiable within the defined scope of the camera family, for example above minimum thresholds for size, visibility and exposure. Pixel size alone does not settle identifiability where information can be combined across frames. Faces and licence plates are scored separately, and each repeated appearance of the same person is scored on its own and tested for linkage (A5).

Reporting detection rates per frame is misleading for continuous visual media. As noted in point 3, a tool missing 1 face in 100 frames at random would already leave about 78% of five-second appearances exposed at least once. Real misses cluster, which is why appearances are measured directly.

Each result states its tested scope (camera family, tool version, settings) and reports a one-sided 95% lower confidence bound, both for the whole test set and for each difficult condition, such as night, rain, crowds, small faces, occlusion and motion blur. Illustrative grades are set out below.

Grade	Lower 95% bound, all appearances	Lower 95% bound, weakest condition	Smallest zero-failure sample
G1	≥ 95%	≥ 90%	59 overall; 29 per condition
G2	≥ 99%	≥ 95%	299 overall; 59 per condition
G3	≥ 99.9%	≥ 99%	2,995 overall; 299 per condition

Grades describe tool performance in a tested scope and do not map to internal use, sharing or publication. Each metric names the failure event it measures, so that coverage, transformation and attack results are reported separately, and a combined grade requires a prespecified, validated rule for combining them. The appearance is an operational unit, whereas the legal record (Glossary) can link several appearances, cameras and recordings. Scope exclusions are predefined and reported, and excluded material counts as unassessed rather than protected. Since repeated people, cameras and reviewers reduce the effective sample size, a prespecified analysis plan covers clustering, subgroup selection and repeated testing.

A grade never excuses a known failure. An exposed appearance found in data to be used or released is corrected, removed or reprocessed with stronger settings, together with linked appearances, annotations or identifiers relating to the same person. Otherwise, the affected material, and any dataset in which it is not treated separately, is handled as containing personal data (paragraph 36).

A4) Test recognition through controlled lineups.

A neutral laboratory runs lineups with consenting volunteers, using both human reviewers and recognition models (face recognition for faces and optical character recognition for licence plates). Two original images of the same person are taken at different moments. Image A, anonymised, is the probe, while image B is placed in a lineup with five distractors matched on attributes the method keeps on purpose, such as age range or gender. Each lineup runs twice, first with the original image A as a control and then with the anonymised image A, with different human reviewers in each round. Control thresholds and exclusions are prespecified and reported, and human and machine results are assessed separately, so that a case a machine can solve is not dropped because humans cannot. Chance success in this design is 1 in 6 for humans and models alike.

The pilot adds several safeguards to this design. Attack-specific baselines, justified limits on attack advantage and confidence requirements are fixed before testing. A result not significantly above chance does not prove equivalence to chance, since 300 lineups with a result exactly at chance still leave an upper bound of 20.6%, and a true rate of 20% would be detected only about 41% of the time. The laboratory uses a varied set of relevant attack methods, including human reviewers, assumes that attackers know the anonymisation method (paragraph 67), and adds exact-source comparisons where an entity may hold the original footage. Cropped faces isolate leakage from the face itself, while a separate whole-scene test restores clothing, location and context.

Because a method can expose a few people clearly while averaging close to chance, the protocol sets enough repeated trials per person to detect concentrated leakage, and reliable identification of any individual triggers investigation or failure, whatever the aggregate result. For licence plates, an anonymised plate is shown with six typed numbers of the same format, one of them real, and tested across the whole sequence, because different characters exposed in different frames can reveal the full plate. Limits are set in the laboratory protocol and justified against the threat context (paragraph 87, Example 20). Where recognition testing is extended beyond consenting laboratory volunteers to field footage of real bystanders, running face matching may constitute processing of biometric data, requiring an Article 6 legal basis and, where Article 9 applies, a condition under Article 9(2).

A5) Test the specific risks of synthetic replacement.

Synthetic replacement keeps video useful for training but brings risks of its own, which the pilot tests within a stated scope rather than through universal promises. Leakage of the source identity into the replacement is tested through recognition attacks against the source (A4) and attempts to recover the source identity. Resemblance to another real person, including the copying or memorising of training faces, is checked by the tool maker once per model version against its own

training data and a defined reference set, rather than through an ongoing comparison with external databases of real people. Generated licence plates follow a documented policy. Where they are random but valid combinations, a coincidental match with a real registration cannot be excluded, but it is unlikely to point to a real vehicle unless the make, model and colour also match, and the policy records how this risk is assessed. Retained original characters are checked across the sequence.

A generated identity that persists across recordings is tested with volunteers filmed on different days and printed plates on a closed site, while contextual linkage through clothing, gait, vehicles and metadata is tested for every method. Attributes kept on purpose, such as gaze or expression, are declared and tested together with their context.

A6) Provide evidence of coverage and transformation integrity.

Coverage needs positive evidence for every method, since a visual check of the output alone cannot reliably show that every face was found and sufficiently transformed. Anonymisation tools can provide this evidence directly by recording which regions they processed, so that audit tools can distinguish processed from unprocessed regions automatically. Users may rely on any combination the laboratory has validated, such as a secure, temporary comparison of input and output, independently inspected transformation masks, processing records tied to outputs and settings, validated machine-readable markers or, where necessary, retained samples under strict governance.

None of these replaces recognition testing. A marker shows that a region was processed, not that every face was found or sufficiently changed, and it may not survive compression or cropping. The approach adds as little exposure as possible, so that it remains compatible with anonymisation at capture and prompt deletion of originals (paragraph 39).

Machine-readable marking under Article 50(2) AI Act is a separate legal question. The Commission's guidance states that Article 50 applies from 2 August 2026, with a limited transition for Article 50(2) marking and detection until 2 December 2026 for systems placed on the market before 2 August 2026. Its guidelines allow a choice of marking techniques and, subject to the applicable scope and exceptions, treat pixelation or blurring of faces as a minor alteration but face replacement as requiring marking. Meeting marking duties is not evidence that identities are protected.

A7) Check each deployment on real-world footage.

Each user checks its own cameras on a random sample of recent footage, using a free toolkit. The full procedure below is intended for large-scale AI training datasets and data shared with other parties, while smaller or purely internal deployments, in particular where originals are deleted promptly after anonymisation (paragraph 39), may rely on a lighter sample check. The user first lists all footage in scope for the period, with documented exclusions and processing failures, and fingerprints each file before the draw, since a fingerprint proves that a file is unchanged but only a complete inventory shows that difficult footage could be drawn. A public random value then selects recordings or camera-time blocks, independently of the detections of the tool, so that anyone can repeat the draw. Sampling by appearance is acceptable only where the completeness of the appearance inventory has itself been established.

Independent annotators examine all relevant appearances in the sampled material, including anything that audit software did not flag, and review the whole data package, such as annotations, timestamps, location data, captions and identifiers. Neither selection nor ground truth depends solely on the detections of the anonymisation tool. The resulting appearance-level rates are reported with confidence bounds per condition, tied to the tool version and settings, and kept as documentation under paragraph 41. Finally, an auditor repeats the draw and re-checks a sample.

Only footage the tool cannot have learned from counts: footage recorded after the tool version was frozen, or independently held historical footage shown not to have been used in training. Effort and frequency remain open questions for the pilot. A candidate design is a yearly check plus targeted revalidation after material changes that could affect privacy effectiveness or the anonymity assessment, such as a new detector or camera domain, changed access arrangements, newly available auxiliary information, significant new attacks or a security incident.

A8) Report utility separately.

Once privacy outcomes are met, results also show what remains useful for the intended task, such as detection and segmentation performance, preservation of pose, motion or gaze where needed, temporal consistency, distortion of annotations, unnecessary redaction, and throughput, latency and cost. Utility is never merged with privacy into one score, so that a useful output cannot offset a privacy failure.

A9) Use context triage to decide on additional safeguards.

A short triage helps decide which additional safeguards and review a deployment needs. It does not decide whether data is anonymous or whether a use is permitted, and legal prohibitions, such as those in Article 5 AI Act, are assessed separately. Distinctive non-facial attributes or sparse scenes call for a whole-scene recognition test and, where appropriate, whole-body treatment. Precise time and place, or links to named records such as badge logs, call for coarsening time and place, restricting linkable metadata and reviewing recipients in context. Repeated appearances across days, as with fixed cameras or fleets on fixed routes, call for a linkage test across recordings.

Sensitive places or people, such as clinics, places of worship, demonstrations, workplaces or children, call for DPIA screening and, where required, a DPIA, together with an enhanced review of transformation, aggregation and disclosure. Wider disclosure, such as more recipients, exports or public release, calls for larger safety margins (paragraph 88) and a broader set of relevant entities.

A10) Compare tools only within camera families.

Scores are compared only within a camera family, so that a street camera is never judged on vehicle footage. A family groups cameras whose images look alike to the software, based on the sensor and colour filter (e.g. standard colour, RCCB as common in vehicle cameras, infrared or thermal), the lens (e.g. standard, wide-angle or fisheye), the viewpoint (e.g. vehicle front, rear or cabin, street pole, indoors or drone), the size of people in the image measured in pixels, and the processing (e.g. HDR or compression settings, frame rate, and whether the tool runs on the device or in the cloud).

The pilot starts with the three or four most common families. Each user describes its setup on a one-page camera card, and where no family fits, the real-world check on its own footage provides scoped deployment evidence, without conferring a transferable laboratory grade.

A11) Ensure neutral governance.

A neutral programme sets the rules, and the companies being tested never hold the test data. A programme board, composed of users such as vehicle makers and cities, research institutes and civil society, with data protection authorities invited as observers, owns the rules, appoints laboratories and publishes results. Alongside it, a technical committee of laboratory experts, users and tool makers, with declared conflicts of interest, advises on test methods but does not decide individual results. The secret test videos are held by neutral laboratories, such as universities, which also run the laboratory tests and decide on results. Tool makers submit their tools and pay published fees, never see the test videos and may appeal results. Users run real-world checks with a free toolkit, and auditors spot-check those checks.

Results can be appealed and corrected through a published procedure, and changes to test criteria are announced with their evidence and a transition period. The pilot would start with a narrow scope in months 0 to 3, covering rules, toolkit and first self-checks for one or two camera setups, followed by a first laboratory round in months 3 to 9 covering three or four camera families, with first results published. From month 9, auditor spot-checks would begin and criteria would be reviewed on a published basis. From year 2, the programme could optionally be formalised as a GDPR code of conduct (Article 40) or certification (Article 42). No new law is needed to start.

References

- EDPB, Guidelines 02/2026 on Anonymisation, version 1.0 for public consultation, adopted 7 July 2026
- EDPB, public consultation page, including the summary “Guidelines at a glance” (October 2026)
- EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models
- European Commission, Guidelines on the transparency obligations for certain AI systems under Article 50 AI Act, C(2026) 5054, 20 July 2026
- European Commission, FAQ: Transparency obligations under Article 50 of the AI Act
- R. McPherson, R. Shokri and V. Shmatikov, Defeating Image Obfuscation with Deep Learning, arXiv:1609.00408 (2016)
- GB/T 44464—2024, General requirements of vehicle data, recommended national standard of China, issued and implemented 23 August 2024: clause 5.6.2 and Annex B
- Euro NCAP, How cars are tested for safety