Comments on EDPB Guidelines 02/2026 on Anonymisation

Public consultation, 8 July to 30 October 2026

Submitted by:
Dr. Juan Emmanuel Delva Benavides
Dr. Jorge Antonio Leos Navarro
Dr. Alejandro Paul García Hernández

Research Group (Cuerpo Académico) UDEG-CA-1236 "Law and Technology", Department of Social and Legal Sciences, University Centre for Economic and Administrative Sciences (CUCEA), University of Guadalajara, Mexico

Contributor type: academic / research institution · Country: Mexico

This contribution reflects the academic views of its authors as members of the research group. It does not represent an institutional position of the University of Guadalajara.

1. Introduction

In 2000, Latanya Sweeney estimated that 87% of the population of the United States was likely to be uniquely identified by three attributes alone: five-digit ZIP code, gender and date of birth. Nearly two decades later, Rocher, Hendrickx and de Montjoye showed that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Over the same period, the label "differential privacy" came to cover deployments of very different strength: the U.S. Census Bureau set a total privacy-loss budget of ε = 19.61 for the 2020 Census redistricting data, while independent researchers found that Apple's implementation on macOS allowed a privacy loss of up to 16 *per day* for some data types, renewed daily. The two figures are not directly comparable, since one concerns a single statistical release under the central model and the other a recurring collection under the local model, and that is precisely the point: the words "differentially private" say little unless the parameters, the unit of protection and the time horizon are stated. Machine learning has added a further layer. In 2021 Carlini and colleagues extracted hundreds of verbatim sequences from the training data of a publicly released language model, including names, telephone numbers and e-mail addresses.

The common thread is the premise at the centre of the draft Guidelines 02/2026: anonymity is not a property of a technique but the outcome of an assessment of what those who may access the data can actually do with it. We welcome the Guidelines as a long-awaited update of Opinion 05/2014 of the Article 29 Working Party. They bring the case law of the Court of Justice of the European Union (CJEU) into the analysis, in particular *EDPS v SRB* (C-413/23 P) on the relative, entity-specific character of personal data (¶6–7, ¶11–15, ¶21–22). They also refine the three criteria of No Record Isolation, No Linkage and No Inference, and they offer a contextual and a simplified approach that controllers can combine (¶45–49). Several choices deserve explicit support: the standard of an insignificant rather than a zero likelihood of identification (¶22); the refusal to treat "lack of motivation" as a factor (¶31); the possibility of rebutting the presumption that people follow the law with concrete evidence (¶33); the recognition that attacks can extract data from supposedly anonymous AI models (¶77, ¶82); the warning that agentic AI will reduce the time and cost of re-identification (¶92–93); and the recommendation to reassess anonymity periodically (¶35).

We write from Mexico, as we did in our contribution to Guidelines 03/2026 on web scraping. Regulators in Latin America read EDPB guidance closely, and data concerning persons in the region is routinely processed by controllers in the Union and by non-EU controllers that serve the European market. Our com-

ments also draw on our published research on the tension between algorithmic transparency and differential privacy (Delva Benavides & Méndez González, 2025), which described how noise that protects identities can also conceal discriminatory patterns (what we called "second-order opacity") and proposed a layered model of access to algorithmic information. The Guidelines are deliberately neutral as to technique, and we agree with that choice. Our proposals seek to make the framework more operational where the stakes are now highest: formal privacy guarantees, noise and synthetic data, AI models and explanation interfaces, the interplay with algorithmic accountability under the Artificial Intelligence Act (AI Act), and identifiers and auxiliary information from outside the Union.

2. Summary of recommendations

- **Formal privacy guarantees (¶67–71, footnotes 43–44):** acknowledge that the comparison between A(D) and A(D′) in ¶71(b) is the test that differential privacy formalises, and set out what a controller relying on it should document: parameters (ε, δ), unit of privacy, trust model, cumulative accounting across releases and over time, values published without noise, and verification of the implementation.
- **Noise and synthetic data (¶64, ¶76, ¶82):** state expressly that neither noise addition nor synthetic data generation makes data anonymous per se, and require documented testing against linkage, attribute inference and membership inference, with attention to outliers.
- **AI models and explanation interfaces (¶77, ¶79, ¶82, footnote 45):** incorporate into the Guidelines the two-pronged test and the attack families of Opinion 28/2024 (¶43 and ¶55), and treat the outputs of query and explanation interfaces as given data, so that anonymisation is never invoked to withhold the explanations required by *Dun & Bradstreet Austria* (C-203/22).
- **Anonymisation and accountability (¶41, Example 14):** ask controllers to document the effect of anonymisation on subgroup-level utility, which is the evidence that Article 4a(1)(a) of the AI Act, inserted by Regulation (EU) 2026/1744, now requires; illustrate tiered access under the contextual approach; and clarify the legal consequences of applying insights from anonymous data to new individuals.
- **Third-country identifiers and auxiliary information (¶24, ¶29(c), ¶33, ¶50, ¶61):** require that assessments consider structured national identifiers that encode quasi-identifiers, the public registries of the jurisdictions concerned and documented leaks of protected registries.
- **Distributed privacy-enhancing technologies (¶91):** extend the reasoning on encryption to federated learning, secure aggregation, secure multi-party computation and knowledge distillation.
- **Editorial corrections:** fix internal cross-references and wording in footnote 44, ¶53, ¶71(c), Example 14, ¶85, ¶89 and the table in ¶97.

3. Detailed comments

3.1 Formal privacy guarantees: making the "neighbouring dataset" test operational (¶67–71, footnotes 42–44)

The treatment of the No Inference criterion is, in our view, the most important contribution of the Guidelines. Paragraph 67 defines a meaningful inference as one that relies on the given data and could not be obtained from general knowledge or from data about the population at large. Paragraph 71(b) and footnote 43 propose to demonstrate this by comparing the outcome of an inference drawn from A(D), the anonymised output of the original dataset D, with the outcome drawn from A(D′), where D′ is the same

dataset without the records of the individual concerned. Footnote 44 then asks whether the given data yields a "significantly different posterior belief". This is, in substance, the definition of differential privacy introduced by Dwork in 2006: a randomised mechanism is ε-differentially private if, for any two datasets that differ only in the data of one individual, the probability of any output changes by at most a factor of e^{ε}. A differentially private mechanism therefore bounds, by design and for every individual at once, the very difference between A(D) and A(D′) that ¶71(b) invites controllers to test case by case. In the terms of footnote 44, with ε = 1 an adversary's odds about any individual can change by at most a factor of about 2.7. Yet the Guidelines do not mention differential privacy, although it is the most widely deployed formal privacy framework in official statistics and in industry.

We do not suggest that differential privacy should become a safe harbour. Its guarantee is only as strong as its parameters, which in practice vary by orders of magnitude, as the figures in our introduction show. It also has a precise scope. It limits what can be learned about a person from the inclusion of their data. It does not prevent inferences drawn from population-level patterns, which, consistently with ¶70 and Examples 13 and 14, the Guidelines rightly treat as not meaningful. Precisely because the Guidelines' own test and the formal definition coincide, however, a brief statement would give controllers and supervisory authorities a shared vocabulary and a measurable basis for the demonstration that ¶71(b)–(c) requires.

We therefore suggest adding a paragraph after ¶71 stating that, where a controller relies on a formal privacy guarantee to show that the No Inference criterion is met, the documentation required by ¶41 should include: (a) the privacy-loss parameters (ε and, where applicable, δ) and an explanation, in plain terms, of what they mean for the posterior belief of an adversary; (b) the unit of privacy (event, record, user or household) and the trust model (central or local); (c) cumulative accounting of the privacy loss across all releases produced from the same original data and over time, consistent with ¶77 on multiple aggregate datasets and ¶35 on periodic reassessment; (d) any statistics published exactly, without noise, since such values fall outside the guarantee; and (e) verification that the implementation actually delivers the claimed guarantee, because an implementation error can void it. Empirical auditing methods for this purpose exist (Jagielski, Ullman & Oprea, 2020; Nasr et al., 2023), as do open-source tools such as Google's DP-Auditorium. The EDPB could also state that larger parameter values call for a stronger justification and a full contextual assessment, and announce further work, for instance with ENISA, on reference ranges for common use cases.

3.2 Noise and synthetic data are techniques, not outcomes (¶64, ¶76, ¶82)

Paragraph 64 correctly warns that linkage remains possible on records that have been intentionally modified, citing the robust de-anonymisation attacks of Narayanan and Shmatikov, and ¶82 mentions synthetic data among the outputs from which specific inferences can be drawn. Practice, however, often treats synthetic data as anonymous by construction. The most systematic evaluation to date points the other way. Stadler, Oprisanu and Troncoso (2022) showed that synthetic data either fails to prevent linkage and attribute inference attacks or, when generated with enough protection to resist them, loses much of its utility, and that the privacy gain for outlier records is especially unpredictable. This is consistent with the Guidelines' own observation that inference is more likely for outliers (¶76).

We suggest stating expressly, in or after ¶82, that adding noise to records or generating synthetic data does not by itself make data anonymous, and that the resulting dataset should be tested against the three criteria. Particular attention should go to membership inference (¶79) and to outliers (¶76), and the documentation should record the success rate of each attack against a baseline that reflects the adversary's prior belief (footnote 44). An example would help: a synthetic dataset of hospital patients that reproduces,

almost unchanged, the record of the only patient with a rare disease in a small municipality, thereby allowing both membership inference and the inference of the diagnosis.

3.3 AI models, query interfaces and explanations (¶77, ¶79, ¶82, footnote 45)

The Guidelines recognise that attacks can extract data from supposedly anonymous AI models (¶77) and that querying or prompting a model can elicit information about a particular individual (¶82). Opinion 28/2024, however, appears only in footnote 45, although it already sets out a precise standard: an AI model can be considered anonymous only if both the likelihood of direct, including probabilistic, extraction of personal data concerning the individuals whose data were used to train it and the likelihood of obtaining such data from queries are insignificant (Opinion 28/2024, ¶43). The same Opinion lists the attacks against which models should be tested: attribute and membership inference, exfiltration, regurgitation of training data, model inversion and reconstruction (¶55). The research record supports this standard. Shokri and colleagues (2017) showed that membership in a training set can be inferred from a model's outputs alone, Fredrikson, Jha and Ristenpart (2015) recovered recognisable facial images from a face recognition model, and Carlini and colleagues (2021) extracted verbatim training data from a language model. We suggest incorporating the two-pronged test and the list of attacks into ¶82, so that the general guidance on anonymisation and the Opinion on AI models apply one and the same standard. Doing so would also clarify that, for trained models, the No Inference criterion will usually be decisive, since No Record Isolation is difficult to apply to model parameters.

A second point concerns explanation interfaces. In *Dun & Bradstreet Austria* (C-203/22, 27 February 2025) the CJEU held that the data subject is entitled to an explanation of the procedure and principles actually applied in an automated decision, and accepted that, for profiling such as credit scoring, it may suffice to show how a change in the person's data would have changed the result, which is in effect a counterfactual explanation (Wachter, Mittelstadt & Russell, 2018). Explanations, however, disclose information about the model and, potentially, about other individuals in its training data. Shokri, Strobel and Zick (2021) showed that model explanations can increase the risk of membership inference and help reconstruct training records. The Guidelines should note that the outputs of query and explanation interfaces offered to data subjects or to the public are themselves given data whose anonymity, with respect to other individuals, must be assessed. They should add that such risks call for safeguards by design (rate limits, control of output granularity, differentially private training) rather than for refusing the explanation. Otherwise, anonymisation risks becoming an argument for withholding the explanations that the Court has now made enforceable.

3.4 Anonymisation should not become "second-order opacity" (¶40–41, ¶47, ¶89–90, Example 14)

In our research we described as "second-order opacity" the situation in which noise or generalisation that protects identities also blurs the correlations needed to detect discrimination through proxy variables: the dataset complies with data protection on its face, yet becomes unusable for auditing bias (Delva Benavides & Méndez González, 2025). The concern is empirical. Bagdasaryan, Poursaeed and Shmatikov (2019) showed that training with differential privacy reduces model accuracy more for under-represented groups than for the majority. The question now also has a legal dimension. Under Article 4a(1)(a) of the AI Act, inserted by Regulation (EU) 2026/1744 (the Digital Omnibus on AI), in force since 27 July 2026, providers of high-risk AI systems may process special categories of personal data for bias detection and correction only where this "cannot be effectively fulfilled by processing other data, including synthetic or anonymised data", and their records of processing must explain why. Every provider relying on Article 4a will therefore

have to show whether anonymised data would have sufficed. The documentation of the anonymisation process required by ¶41 is the natural place for that evidence.

We suggest three additions. First, ¶41 could state that, where anonymised data is intended for testing accuracy or detecting bias, the documentation should record which statistical properties were preserved and the measured effect of the anonymisation on subgroup-level utility, for example error rates for groups defined by protected characteristics. Second, an example under the contextual approach (¶47, ¶89–90) could illustrate tiered access to the same original data: (i) personal data held under strict access controls for supervisory authorities and accredited auditors, where the GDPR applies in full and no anonymity is claimed; (ii) a secure research environment, assessed contextually; and (iii) public statistics or model documentation protected by formal guarantees and assessed under the simplified approach. This is the layered model we proposed in our research, and it would show that the relative concept of personal data confirmed in *EDPS v SRB* can serve accountability, not only data sharing. Third, Example 14 could add that, while the historical loan dataset may pass the No Inference criterion, applying the resulting insight to the new applicant is a processing of that applicant's personal data. That processing is subject to the GDPR, including Article 22 and the explanation duties clarified in C-203/22, and credit scoring is a high-risk use under Annex III, point 5(b), of the AI Act. Passing the criteria settles the status of the given data, not the legality of what is done with its insights.

3.5 Third-country identifiers, public registries and evidence of past leaks (¶23–25, ¶29(c), ¶30, ¶33, ¶50, ¶61, ¶89)

The examples in the Guidelines rely on European attributes such as postcodes, sex and date of birth. Datasets processed in or from the Union, however, increasingly concern persons elsewhere, and their quasi-identifiers follow other patterns. Mexico offers a clear illustration. The Unique Population Registry Code (CURP) is an 18-character code that encodes letters of the person's surnames and first name, the date of birth, sex and state of birth. Under the reform of the General Population Act published on 16 July 2025, the CURP, in a version that incorporates fingerprints and a photograph, becomes the mandatory national identity document, and every public or private entity must request it for its procedures and services (Articles 91 Bis and 91 Sexies). The Federal Taxpayer Registry code (RFC) similarly encodes name initials and the date of birth. Two consequences follow. A dataset that retains the CURP or the RFC, even in truncated form, retains a bundle of quasi-identifiers (¶24). And a hash of the CURP is a pseudonym that can easily be reverse-engineered within the meaning of ¶50: anyone who knows a person's name, date of birth, sex and state of birth needs to compute and compare at most a few dozen candidate codes.

The presumption that the law will be respected (¶32–33) also needs to be tested against local evidence. In April 2016 a copy of Mexico's voter list, containing the personal data of about 93.4 million voters, was found openly accessible on a commercial cloud server. The political party that had received the list acknowledged the upload; the National Electoral Institute fined it approximately MXN 34 million, a sanction confirmed by the Federal Electoral Tribunal in 2018. This is exactly the kind of evidence that the third and fourth indents of ¶33 describe: a registry protected by law whose protection did not prevent its dissemination. A contextual assessment of a dataset concerning residents of Mexico should therefore not assume that such registries are beyond the reach of a relevant entity. Judicial bulletins that list the names of parties by case number, and public versions of court decisions whose redaction practices vary from court to court, raise similar questions.

We suggest adding to ¶29(c) and ¶61 that the additional information to be considered includes the public registries, the structure of national identifiers and the documented leaks of the jurisdictions where data

subjects live and where recipients operate. This matters especially for transfers under Chapter V of the GDPR and for non-EU controllers subject to Article 3(2). Paragraph 50 could also mention hashed structured identifiers among the pseudonyms that can easily be reverse-engineered, with an example drawn from a national identity code.

3.6 Distributed privacy-enhancing technologies (¶91)

Paragraph 91 explains that encryption can support anonymisation by limiting access to intelligible data, but differs from it because it is reversible by design. The same reasoning would usefully extend to other privacy-enhancing technologies now common in AI development. In federated learning (McMahan et al., 2017), raw data remain on users' devices, but individual model updates can leak training data (Zhu, Liu & Han, 2019) and should generally be treated as personal data. Secure aggregation (Bonawitz et al., 2017) and secure multi-party computation restrict access to individual inputs and can therefore support, but not by themselves constitute, anonymisation. The aggregated or distilled model that results (Hinton, Vinyals & Dean, 2015) should then be assessed under ¶82 and the test of Opinion 28/2024. A short paragraph to this effect would prevent the architecture of a system from being mistaken for the anonymity of its outputs.

3.7 Editorial corrections

In the course of our reading we noted a few cross-references and wordings that could be corrected in the final version. They are listed at the end of the table in section 4.

4. Drafting suggestions

Paragraph	Issue	Suggested change
New ¶ after ¶71	Formal privacy guarantees are not addressed, although ¶71(b) and footnote 43 describe the test they formalise	Recognise differential privacy as a formalisation of the A(D)/A(D′) comparison; list the documentation items (ε, δ, unit, trust model, cumulative accounting, values published without noise, verification of the implementation).
¶64, ¶82	Noise and synthetic data treated as anonymous by construction in practice	State that neither is anonymising per se; require testing with documented attack success rates against a prior-belief baseline; add a synthetic-data example with an outlier record.
¶82, fn 45	Two instruments, potentially two standards for AI models	Incorporate the two-pronged test of Opinion 28/2024 (¶43) and its list of attacks (¶55).
New ¶ after ¶82	Query and explanation interfaces	Treat their outputs as given data; require safeguards by design so that anonymisation is not invoked to withhold explanations required by C-203/22.
¶41	Documentation silent on utility for bias detection	Record the preserved statistical properties and the effect on subgroup-level utility, to support the necessity assessment under Article 4a(1)(a) AI Act.
¶47, ¶89–90	No example of tiered access	Add an example: authorities and auditors (personal data); secure research environment (contextual approach); public outputs (simplified approach).
Example 14	Legal position of the new applicant	Add that applying the insight is a processing of the applicant's personal data (Article 22 GDPR; C-203/22; Annex III, point 5(b), AI Act).

Paragraph	Issue	Suggested change
¶29(c), ¶61, ¶33	Auxiliary information calibrated to European contexts	Include third-country public registries, identifier structures and documented leaks; recall Chapter V and Article 3(2) GDPR.
¶50	Hashed structured identifiers	List them among easily reversible pseudonyms, with an example based on a national identity code.
¶91	Only encryption is addressed	Extend to federated learning, secure aggregation, secure multi-party computation and knowledge distillation.
Footnote 44	Incorrect cross-references	"see footnote 40" → "see footnote 42"; "A(D′) in footnote 41" → "A(D′) in footnote 43".
¶97 (table)	Typographical error	"See paras. 98 to 00" → "See paras. 98 to 100".
¶71(c)	Terminology	"rebuttal presumption" → "rebuttable presumption".
¶85	Wording	"whose data is included in the record" → "whose data is included in the given data".
¶89, ¶53, Ex. 14	Wording	"a mean reasonably likely" → "a means reasonably likely"; "reintroduced sub-section 3.5" → "reintroduced in sub-section 3.5"; "uses this insight to a new loan applicant" → "applies this insight to a new loan applicant".

5. Concluding remarks

The draft Guidelines replace a twelve-year-old benchmark with a framework that is legally sound and, in its treatment of inference, conceptually ambitious. Their success will depend on how far controllers can turn that framework into evidence a supervisory authority can verify. Our proposals aim at that. The EDPB should recognise that its own test for meaningful inference is the one that differential privacy formalises, and ask for the parameters that make such claims checkable. It should state that noise and synthetic data are techniques, not outcomes, and align the standard for AI models with Opinion 28/2024, including the interfaces through which models are queried and explained. It should ensure that anonymisation does not become a new form of opacity, now that Article 4a of the AI Act makes the utility of anonymised data for bias detection a legal question. And it should require assessments to look at the identifiers, registries and leaks of the places where data subjects actually live. Expectations of this kind would protect data subjects on both sides of the Atlantic and would give the authorities and legislators of Latin America, who are re-building their own institutions, a reference they can adopt with confidence. We remain available to the EDPB for any further exchange.

References

EDPB, *Guidelines 02/2026 on Anonymisation*, version 1.0, adopted 7 July 2026 (version for public consultation).

EDPB, *Guidelines 03/2026 on web scraping in the context of generative AI*, version 1.0, adopted 7 July 2026.

EDPB, *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*, adopted 17 December 2024.

Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques* (WP216), adopted 10 April 2014.

CJEU, C-413/23 P, *EDPS v SRB*, judgment of 4 September 2025; C-203/22, *Dun & Bradstreet Austria*, judgment of 27 February 2025.

Regulation (EU) 2016/679 (GDPR); Regulation (EU) 2024/1689 (AI Act), as amended by Regulation (EU) 2026/1744 (Digital Omnibus on AI).

Mexico: Decree amending the General Population Act (Ley General de Población), Diario Oficial de la Federación, 16 July 2025; Federal Law on the Protection of Personal Data Held by Private Parties, Diario Oficial de la Federación, 20 March 2025.

U.S. Census Bureau, "Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results", press release, 9 June 2021.

Bagdasaryan, E., Poursaeed, O. & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems* 32 (NeurIPS 2019).

Bonawitz, K. et al. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191.

Carlini, N. et al. (2021). Extracting training data from large language models. *30th USENIX Security Symposium*, 2633–2650.

Delva Benavides, J. E. & Méndez González, V. V. (2025). Algoritmos opacos y privacidad diferencial: ¿puede haber transparencia sin comprometer la protección? *Forseti. Revista de Derecho*, 14(22), 227–245. https://doi.org/10.21678/forseti.v14i22.2834

Dwork, C. (2006). Differential privacy. *33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, LNCS 4052, 1–12.

Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.

Fredrikson, M., Jha, S. & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–1333.

Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv:1503.02531.

Jagielski, M., Ullman, J. & Oprea, A. (2020). Auditing differentially private machine learning: How private is private SGD? *Advances in Neural Information Processing Systems* 33 (NeurIPS 2020).

McMahan, B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 54, 1273–1282.

Narayanan, A. & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy*, 111–125.

Nasr, M. et al. (2023). Tight auditing of differentially private machine learning. *32nd USENIX Security Symposium*.

Rocher, L., Hendrickx, J. M. & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10, 3069.

Shokri, R., Stronati, M., Song, C. & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy*, 3–18.

Shokri, R., Strobel, M. & Zick, Y. (2021). On the privacy risks of model explanations. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*.

Stadler, T., Oprisanu, B. & Troncoso, C. (2022). Synthetic data – anonymisation groundhog day. *31st USENIX Security Symposium*.

Sweeney, L. (2000). *Simple demographics often identify people uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3.

Tang, J., Korolova, A., Bai, X., Wang, X. & Wang, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. arXiv:1709.02753.

Wachter, S., Mittelstadt, B. & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887.

Zhu, L., Liu, Z. & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems* 32 (NeurIPS 2019).