Adaptive DNN Partitioning for Low-Latency Privacy-Preserving Edge Computing

Collette Elwood
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
colletteelwood05@ku.edu

Marsudi Padmasari
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
marsudi2002@unr.edu

Feline Uphaus
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
felinework@uc.edu

Abstract

Edge computing has become a central strategy for reducing the latency, bandwidth, and privacy risks associated with cloud-centric deep neural network inference. However, moving computation closer to data sources does not automatically resolve the tension between responsiveness and confidentiality. Adaptive deep neural network partitioning offers a system-level mechanism for distributing model layers across end devices, edge servers, and cloud infrastructure according to changing network conditions, compute capacity, energy constraints, and privacy requirements. This paper examines adaptive partitioning as a socio-technical architecture rather than a purely algorithmic optimization problem. It analyzes structural trade-offs among latency, accuracy, communication overhead, hardware heterogeneity, trust boundaries, and long-term operational sustainability. The discussion emphasizes privacy-preserving inference through differential privacy, federated learning, trusted execution environments, and secure enclaves, while also considering governance, fairness, robustness, and policy implications. Case illustrations from healthcare, manufacturing, autonomous mobility, and smart cities demonstrate that partitioning decisions are embedded in institutional and infrastructural contexts. The paper argues that low-latency privacy-preserving edge computing requires adaptive orchestration, transparent accountability, and resilient deployment models that can accommodate device churn, adversarial threats, regulatory diversity, and uneven access to computational resources. The conclusion outlines open challenges in cross-layer optimization, verifiable trust, sustainable edge infrastructure, and equitable participation in edge artificial intelligence ecosystems.

Keywords

adaptive DNN partitioning; edge computing; privacy-preserving inference; low latency; trusted execution; federated learning; socio-technical governance 1 Introduction Edge computing has emerged as a response to the growing mismatch between centralized cloud intelligence and the latency-sensitive, data-intensive, and privacy-critical demands of contemporary applications [1]. The vision of pushing computation toward the network periphery is not merely a matter of geographic proximity. It reflects a broader reconfiguration

of infrastructure in which sensing, inference, storage, and control are distributed across heterogeneous devices and service domains [2]. Mobile edge computing further extends this model by integrating cellular networks, access points, and localized compute nodes into the service path, thereby enabling applications that cannot tolerate distant round trips or continuous backhaul consumption [3]. Deep neural networks are among the most demanding workloads in this environment, because they combine high computational intensity with large intermediate representations and complex data dependencies. The central problem addressed in this paper is how to partition deep neural network inference across end devices, edge servers, and cloud resources in a manner that simultaneously reduces latency and preserves privacy. Static partitioning is often inadequate because network conditions fluctuate, device capabilities vary, privacy requirements differ across jurisdictions and sectors, and workloads change over time. A partition that performs well for a smartphone on a high-bandwidth network may be inappropriate for an industrial sensor on a constrained wireless link or for a medical device subject to strict confidentiality rules. Adaptive partitioning therefore becomes a control problem that spans model architecture, runtime scheduling, communication protocols, hardware acceleration, and governance. The challenge is not only to find an efficient split point but also to determine who may observe which intermediate representations, under what legal and institutional conditions, and with what accountability mechanisms. This paper contributes a system-oriented analysis of adaptive deep neural network partitioning for low-latency privacy-preserving edge computing. It emphasizes architecture, structural trade-offs, infrastructure, deployment, sustainability, robustness, fairness, and policy. Rather than treating privacy as a post hoc encryption layer, the analysis treats trust boundaries as first-class design constraints that shape where computation can occur and how intermediate data can move. The discussion also considers the limits of decentralization, because edge infrastructure itself depends on energy, maintenance, supply chains, standards, and public or private investment. The remainder of the paper reviews related work, examines partitioning architecture, analyzes privacy and trust, discusses adaptive control and robustness, considers governance and sustainability, presents evaluation and case illustrations, and concludes with open challenges. 2 Background and Related Work Research on collaborative inference between cloud and mobile edge systems established early evidence that neural network computation can be divided across resource domains to improve responsiveness and reduce data transmission [4]. Distributed deep neural networks extended this idea by executing different portions of a model across cloud, edge, and end devices, thereby creating opportunities for hierarchical inference and early exits [5]. Surveys of deep learning in mobile and wireless networking have further documented the breadth of deployment scenarios, from latency-critical vision tasks to bandwidth-constrained sensing systems [6]. These works collectively show that partitioning is not a single technique but a family of architectural choices involving layer placement, feature compression, model segmentation, and runtime coordination. Recent work has explored model adaptivity for efficient private inference in edge computing, demonstrating that partitioning strategies can be adjusted to balance privacy protection against computational and communication costs [7]. This line of inquiry is important because privacy-preserving mechanisms often introduce overhead that interacts with latency and energy budgets. Complementary research on execution time optimization for mobile and embedded devices has shown that neural network performance is highly sensitive to memory access patterns, operator scheduling, and hardware-specific constraints [8]. Dynamic adaptive DNN surgery has also illustrated how models can be restructured during inference to accommodate edge conditions and accelerate execution [9]. Edge artificial intelligence frameworks have examined on-demand acceleration,

where computation is scheduled according to network and server availability [10]. On-device machine learning research has highlighted algorithmic and learning-theoretic perspectives on resource-constrained inference and training [11]. Large-scale industrial studies of inference at the edge have provided operational evidence about the diversity of workloads, device classes, and deployment constraints in real systems [12]. Privacy-preserving inference draws on several complementary traditions. Differential privacy provides formal noise-based guarantees for statistical disclosure control, although the utility and latency implications require careful calibration in interactive systems [13]. Federated learning enables decentralized model training without centralizing raw data, but it introduces challenges related to communication efficiency, client heterogeneity, and adversarial participation [14]. Broader analyses of federated learning have identified open problems in privacy, robustness, fairness, and system design [15]. Trusted execution environments, such as secure enclaves, offer hardware-assisted isolation for sensitive computation, but they depend on complex attestation and threat models [16]. Security and privacy challenges in industrial Internet of Things deployments further illustrate how device longevity, physical exposure, and safety requirements complicate conventional assumptions [17]. Surveys of mobile edge computing, fog, and related paradigms have documented security threats and challenges across infrastructure, orchestration, and application layers [18]. Together, these bodies of work frame adaptive partitioning as a problem that must be evaluated across technical, institutional, and normative dimensions. 3 System Architecture and Partitioning Trade-offs Adaptive DNN partitioning can be understood as a layered architecture spanning sensing endpoints, edge compute nodes, regional aggregation points, and cloud data centers. Each layer possesses different latency, bandwidth, energy, storage, and trust characteristics. End devices often provide immediate access to raw data and benefit from low transmission delay, but they may lack sufficient memory or acceleration hardware for large models. Edge servers can provide stronger compute while remaining close enough to meet strict latency budgets, yet they may serve multiple tenants and therefore complicate isolation and data governance. Cloud infrastructure offers elastic capacity and sophisticated model management, but it introduces wide-area communication delays and may be subject to different legal regimes. The partition point determines which layer observes raw inputs, intermediate features, and final outputs, making it a privacy-relevant decision as much as a performance decision [4], [7]. The trade-offs are multidimensional. Moving more layers to the edge can reduce communication and improve responsiveness, but it may increase energy consumption at constrained nodes and expand the attack surface of edge infrastructure. Sending compressed intermediate features to the cloud can reduce bandwidth relative to raw data, but intermediate representations may still leak sensitive information through reconstruction or inference attacks. Early-exit models can reduce average latency by allowing simple inputs to be classified near the edge, but they may produce uneven accuracy across populations or conditions. Model splitting can improve parallelism, yet it can also create synchronization overhead and complicate fault tolerance. These tensions mean that partitioning policies cannot be evaluated solely by average latency. They must also consider tail latency, accuracy variance, privacy leakage, energy use, maintenance burden, and operability under degraded conditions [8], [9]. Architecturally, adaptive partitioning requires a control plane that monitors network conditions, device state, workload characteristics, and policy constraints. This control plane may reside at the edge or in a distributed orchestrator, and it must coordinate model versions, feature encodings, cryptographic protections, and fallback modes. In some deployments, the partition point can be changed per inference request; in others, it must remain stable for safety or certification reasons. For example, an autonomous vehicle may require predictable execution paths for

safety certification, while a smart camera may tolerate dynamic reconfiguration to save bandwidth. Similarly, a hospital edge node may need auditable computation paths that satisfy regulatory review, whereas a consumer application may prioritize responsiveness and battery life. The architecture therefore embeds assumptions about acceptable risk, institutional trust, and operational control [10], [12]. A further structural issue is heterogeneity. Edge environments include graphics processors, neural accelerators, digital signal processors, field-programmable gate arrays, and general-purpose processors with different memory hierarchies and software stacks. Adaptive partitioning must account for these differences without assuming a uniform abstraction layer. Compiler techniques, runtime scheduling, and model compression interact with partition selection. A partition that is optimal for one hardware target may be inefficient for another, and the cost of recompilation or model conversion can be significant. This heterogeneity also affects fairness, because users with older or less capable devices may experience different accuracy or privacy protections than users with newer hardware. Responsible design therefore requires explicit consideration of performance envelopes and service-level commitments across device classes. 4 Privacy Preservation and Trust Boundaries Privacy in edge inference is not a binary property. It is a spectrum of guarantees involving data minimization, access control, cryptographic protection, statistical noise, hardware isolation, and contractual governance. Differential privacy can limit the influence of any single data record on model outputs, but it may reduce accuracy and introduce nondeterminism that complicates real-time control [13]. Federated learning can keep raw data local during training, yet model updates may still reveal information, and malicious clients may poison the global model [14], [15]. Trusted execution environments can isolate computation from untrusted operating systems and cloud operators, but they rely on hardware vendors, attestation services, and microarchitectural assumptions that may not hold in all environments [16]. Adaptive partitioning must therefore combine mechanisms rather than depend on any single privacy technology. Trust boundaries are central to the design. If an intermediate feature representation crosses from a trusted device to a semi-trusted edge server, the system must decide whether the representation can be encrypted, perturbed, compressed, or validated. If the edge server is shared among multiple tenants, isolation mechanisms must prevent cross-tenant inference and data remanence. If the cloud is involved, jurisdiction, subpoena exposure, and insider risk become relevant. These concerns are especially acute in industrial and medical settings, where safety, intellectual property, and personal data may coexist in the same inference pipeline [17], [18]. Adaptive partitioning can reduce exposure by keeping sensitive layers local, but it can also create new exposure if intermediate features are more informative than expected or if partition metadata reveals operational patterns. The privacy-latency trade-off is therefore not a simple curve. Some privacy mechanisms add computation and communication overhead, while others change the partition structure itself. For example, secure enclaves may require specific hardware support at the edge, limiting where computation can be placed. Homomorphic encryption may protect data but impose heavy latency costs for large neural networks, making it suitable only for selected operations or low-throughput settings. Secure multi-party computation can distribute trust but requires coordination and network assumptions that may be difficult in mobile environments. A practical adaptive system may use a portfolio of modes: fully local inference for the most sensitive requests, enclave-protected edge inference for moderate sensitivity, and partitioned cloud inference for non-sensitive high-throughput tasks. The required policy engine must map data classifications and user consent to permissible partitions [7], [16]. Privacy also has temporal and relational dimensions. Data that is not sensitive in isolation may become sensitive when linked with other streams, location traces, or inferred attributes.

Edge systems often process continuous sensor data, which increases the risk of re-identification and behavioral profiling. Governance must therefore address purpose limitation, retention limits, auditability, and revocation. Technical mechanisms alone cannot resolve these issues. They must be embedded in institutional arrangements that define who can access logs, how models are audited, and how harms are remedied. This is particularly important when edge inference informs decisions about employment, insurance, policing, or access to public services. Adaptive partitioning should not become a mechanism for obscuring accountability across multiple operators. 5 Adaptive Control, Robustness, and Resource Management Adaptive control for DNN partitioning involves continuous sensing of runtime conditions and policy constraints. The controller may observe wireless link quality, round-trip time, server queue depth, device battery level, thermal state, memory pressure, and model confidence. It may also receive policy signals from a governance layer indicating that certain data types must remain local or that certain edge nodes are temporarily untrusted. The control objective is multi-objective: minimize latency, maintain accuracy, respect privacy, conserve energy, and preserve stability. Because these objectives can conflict, the system needs explicit prioritization rules and fallback behaviors. A low-latency request may tolerate lower accuracy, while a safety-critical request may require conservative local execution even if latency increases [8], [10]. Robustness is essential because edge environments are failure-prone. Wireless links fluctuate, devices move, edge nodes overheat, containers restart, and cloud services experience outages. An adaptive partitioner must detect degradation and reconfigure without violating privacy or safety constraints. It must also resist adversarial manipulation. An attacker might spoof network conditions to force sensitive computation onto an untrusted node, inject corrupted feature representations, or overwhelm the controller with false telemetry. Defenses include authenticated telemetry, policy enforcement at multiple layers, redundancy, and conservative defaults. The system should fail safe by reverting to local inference or a reduced-capability mode when trust cannot be established [9], [17]. Resource management intersects with sustainability. Edge devices are often battery-powered or energy-constrained, and edge data centers consume electricity for computation and cooling. Partitioning decisions that reduce network transmission may increase local energy consumption, while cloud offloading may shift energy use to regions with different carbon intensity. A comprehensive approach would consider energy source, thermal limits, device lifetime, and electronic waste. Sustainability is not merely an environmental add-on; it affects total cost of ownership and long-term service viability. If adaptive partitioning accelerates hardware replacement or requires specialized accelerators that are difficult to recycle, its benefits may be offset by material and logistical burdens. These considerations argue for lifecycle-aware optimization rather than narrow latency minimization [2], [12]. Fairness and robustness are also linked. If adaptive systems allocate more accurate or more private inference to users with better devices or faster connections, they may reproduce existing inequalities. Edge infrastructure is unevenly distributed, and rural or low-income areas may have fewer trusted edge nodes. A fair design would define minimum service guarantees, transparently communicate degradation modes, and provide avenues for redress. It would also consider language, disability, and cultural context in model evaluation. Adaptive partitioning can support fairness by enabling local processing that reduces dependence on distant cloud providers, but it can also fragment accountability. Governance must therefore accompany technical adaptation with measurable commitments and independent oversight [15], [18]. 6 Deployment, Governance, Sustainability, and Fairness Deployment of adaptive DNN partitioning is a complex undertaking involving standards, procurement, interoperability, and operational expertise. Organizations must decide whether to build private edge infrastructure,

rely on telecommunications providers, or use public cloud extensions. Each choice affects control over data, cost structure, latency guarantees, and regulatory compliance. In healthcare, for example, edge inference may support real-time monitoring while keeping patient data within a hospital network, but remote maintenance and software updates must still be governed. In manufacturing, predictive maintenance may require low-latency inference near machines, yet operational technology networks have long lifecycles and strict safety requirements. In smart cities, cameras and sensors raise surveillance concerns that cannot be resolved by technical partitioning alone [19], [20]. Governance must address accountability across distributed components. When an inference error occurs, it may be difficult to determine whether the fault lies in the model, the partition decision, the network, the edge operator, or the cloud provider. Audit trails should record partition choices, policy versions, model versions, and privacy mechanisms without exposing sensitive data. Regulatory frameworks may require explanations, data protection impact assessments, and human oversight. Policy makers should consider whether adaptive partitioning creates new forms of opacity, particularly when intermediate features cross organizational boundaries. The ethics of algorithms literature emphasizes that transparency, fairness, and responsibility are not merely technical properties but social and institutional achievements [20]. Surveillance capitalism critiques further warn that edge intelligence can extend data extraction into everyday life, making governance a matter of power as well as compliance [19]. Sustainability requires attention to both environmental and institutional dimensions. Environmental sustainability includes energy efficiency, carbon-aware scheduling, hardware longevity, and circular economy practices. Institutional sustainability includes funding models, skills, maintenance, and community participation. A city may deploy edge nodes for traffic safety, but if maintenance is unfunded, the system may degrade unevenly. A rural clinic may benefit from local inference, but without technical support and reliable power, the system may fail. Adaptive partitioning should therefore be designed for repairability, observability, and graceful degradation. It should also avoid lock-in to proprietary accelerators or cloud services that make long-term operation dependent on a single vendor. These concerns connect technical architecture to public policy and infrastructure planning. Fairness in deployment also involves participation. Communities affected by edge inference should have a role in decisions about data collection, model use, and redress. This is especially important when systems are deployed in public spaces or essential services. Privacy-preserving partitioning can reduce some risks, but it does not by itself ensure democratic legitimacy. Governance should include transparency reports, independent audits, complaint mechanisms, and clear limits on secondary use. In cross-border deployments, legal differences in privacy, labor, and data localization must be reconciled without creating a race to the lowest protection. Adaptive systems should therefore be designed with configurable policy constraints that reflect local norms and enforceable rights. 7 Evaluation and Case Illustrations Evaluating adaptive DNN partitioning requires more than benchmark accuracy and average latency. A robust evaluation should measure tail latency, throughput under load, energy per inference, memory footprint, communication volume, privacy leakage under attack models, and recovery time after failure. It should also assess fairness across device classes, user groups, and network conditions. Experimental testbeds should include realistic wireless variability, multi-tenant edge servers, and adversarial scenarios. Simulation alone may miss hardware-specific bottlenecks and operational friction. Field deployments provide valuable evidence but raise ethical and logistical challenges, particularly when sensitive data or safety-critical systems are involved. Mixed-method evaluation combining controlled experiments, simulation, and longitudinal observation is therefore appropriate [7], [10]. Several case domains illustrate the interplay of

latency and privacy. In healthcare, wearable and bedside devices may perform local feature extraction, while an edge server runs diagnostic models under strict access controls. Adaptive partitioning could keep sensitive biometric data local and send only encrypted or perturbed summaries to a clinical cloud for longitudinal analysis. However, regulatory requirements, auditability, and clinician trust shape which partitions are acceptable. In autonomous mobility, perception and planning require deterministic low latency, and partitioning across vehicles, roadside units, and cloud services must account for safety certification and liability. In manufacturing, predictive maintenance may benefit from edge inference that avoids exposing proprietary process data, but legacy equipment and operational technology constraints limit rapid reconfiguration [12], [18]. Smart cities combine public infrastructure, private vendors, and diverse populations. Adaptive partitioning could support traffic management, environmental monitoring, and public safety, but it also raises surveillance, equity, and data retention concerns. A privacy-preserving design might process video locally and transmit only anonymized counts or events. Yet even aggregated data can be sensitive when linked with other sources. Governance must therefore define acceptable purposes and oversight. In each case, the partition point is not merely an engineering parameter; it reflects legal requirements, institutional trust, and public expectations. Case illustrations demonstrate that successful systems align technical adaptation with organizational capacity and accountability structures. 8 Open Challenges and Future Directions Several open challenges remain. First, cross-layer optimization must integrate model architecture, hardware scheduling, network dynamics, privacy mechanisms, and policy constraints without creating brittle or opaque control logic. Second, verifiable trust is difficult when hardware attestation, software supply chains, and human operators are all potential weak points. Third, sustainable edge infrastructure requires new metrics and incentives that account for energy, maintenance, and lifecycle impacts. Fourth, fairness across devices, regions, and populations needs measurable standards and inclusive design processes. Fifth, regulatory interoperability must be developed so that adaptive systems can operate across jurisdictions without undermining rights. These challenges are interconnected and cannot be solved by any single discipline [15], [16], [20]. Future research should explore policy-aware orchestration that treats privacy constraints as executable rules. It should investigate hybrid cryptographic and statistical protections that are practical for real-time inference. It should develop robust federated and split-learning methods that tolerate malicious participants and heterogeneous hardware. It should also examine how edge infrastructure can be governed as a public interest resource rather than only a commercial service. Finally, researchers should study failure modes in the wild, including outages, cyberattacks, model drift, and community resistance. Adaptive DNN partitioning is promising, but its legitimacy depends on whether it improves not only latency and confidentiality but also accountability, sustainability, and equitable access [7], [18].

Conclusion

Adaptive DNN partitioning offers a powerful architectural approach for low-latency privacy-preserving edge computing. By distributing inference across end devices, edge servers, and cloud resources according to runtime conditions and policy constraints, it can reduce communication overhead, respect trust boundaries, and accommodate heterogeneous infrastructure. However, the benefits are conditional. Partitioning introduces complex trade-offs among latency, accuracy, energy, robustness, fairness, and governance. Privacy-preserving mechanisms such as differential privacy, federated learning, and trusted execution environments can strengthen protection, but they must be integrated with system-level orchestration and institutional accountability. Deployment in healthcare, mobility,

manufacturing, and smart cities shows that technical choices are embedded in legal, social, and infrastructural contexts. Sustainable and fair edge intelligence therefore requires more than efficient split points. It requires transparent governance, lifecycle-aware infrastructure, robust fallback modes, and inclusive participation. The future of adaptive partitioning lies in designing systems that are not only fast and private but also legitimate, resilient, and equitable.

References

1. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9

2. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

3. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322–2358. https://doi.org/10.1109/COMST.2017.2745201

4. Kang, Y., Hauswald, J., Gao, C., Rovinski, A., Mudge, T., Mars, J., & Tang, L. (2017). Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. ACM SIGARCH Computer Architecture News, 45(1), 615–629. https://doi.org/10.1145/3093337.3037698

5. Teerapittayanon, S., McDanel, B., & Kung, H. T. (2017). Distributed deep neural networks over the cloud, the edge and end devices. In 2017 IEEE 37th International Conference on Distributed Computing Systems (pp. 328–339). IEEE. https://doi.org/10.1109/ICDCS.2017.226

6. Zhang, C., Patras, P., & Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. IEEE Communications Surveys & Tutorials, 21(3), 2224–2287. https://doi.org/10.1109/COMST.2019.2904897

7. Zhou, T., Zhao, J., Luo, Y., Xie, X., Wen, W., Ding, C., & Xu, X. (2024, October). Adapi: Facilitating dnn model adaptivity for efficient private inference in edge computing. In Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design (pp. 1-9).

8. Yao, S., Zhao, Y., Shao, H., Liu, C., Liu, D., Su, L., & Abdelzaher, T. (2019). FastDeepIoT: Towards understanding and optimizing neural network execution time on mobile and embedded devices. In Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems (pp. 278–291). ACM. https://doi.org/10.1145/3274783.3274840

9. Hu, C., Bao, W., Wang, D., & Liu, F. (2019). Dynamic adaptive DNN surgery for inference acceleration on the edge. In IEEE INFOCOM 2019 - IEEE Conference on Computer Communications (pp. 1423–1431). IEEE. https://doi.org/10.1109/INFOCOM.2019.8737614

10. Li, E., Zeng, L., Zhou, Z., & Chen, X. (2019). Edge AI: On-demand accelerating deep neural network inference via edge computing. IEEE Transactions on Wireless Communications, 19(1), 447–457. https://doi.org/10.1109/TWC.2019.2946140

11. Dhar, S., Guo, J., Liu, J., Tripathi, S., Kurup, U., & Shah, M. (2019). On-device machine learning: An algorithms and learning theory perspective. arXiv. https://arxiv.org/abs/1911.00623

12. Wu, C.-J., Brooks, D., Chen, K., Chen, D., Choudhury, S., Dukhan, M., Hazelwood, K., Isaac, E., Jia, Y., Jia, Z., Leyvand, T., Lu, H., Lu, Y., Qiao, L., Qiao, L., Reddi, V. J., Reagen, B., Spisak, J., Sun, F., ... Zhang, P. (2019). Machine learning at Facebook: Understanding inference at the edge. In 2019 IEEE International Symposium on High Performance Computer Architecture (pp. 331–344). IEEE. https://doi.org/10.1109/HPCA.2019.00048

13. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318). ACM. https://doi.org/10.1145/2976749.2978318

14. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273–1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html

15. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210. https://doi.org/10.1561/2200000083

16. Costan, V., & Devadas, S. (2016). Intel SGX explained. IACR Cryptology ePrint Archive, 2016, 86. https://eprint.iacr.org/2016/086

17. Sadeghi, A.-R., Wachsmann, C., & Waidner, M. (2015). Security and privacy challenges in industrial Internet of Things. In Proceedings of the 52nd Annual Design Automation Conference (pp. 1–6). ACM. https://doi.org/10.1145/2744769.2747942

18. Roman, R., Lopez, J., & Mambo, M. (2018). Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. Future Generation Computer Systems, 78, 680–698. https://doi.org/10.1016/j.future.2016.11.009

19. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.

20. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679