Machine learning systems in the IoT
Trustworthiness trade-offs for edge intelligence

Toussaint, Wiebke; Ding, Aaron Yi

DOI
10.1109/CogMI50398.2020.00030

Publication date
2020

Document Version
Accepted author manuscript

Published in
Proceedings - 2020 IEEE 2nd International Conference on Cognitive Machine Intelligence, CogMI 2020

Accepted Author Manuscript. Link to published article (IEEE): https://doi.org/10.1109/CogMI50398.2020.00030

Machine Learning Systems in the IoT: Trustworthiness Trade-offs for Edge Intelligence

Wiebke Toussaint
Engineering Systems and Services
Delft University of Technology
Delft, Netherlands
w.toussaint@tudelft.nl

Aaron Yi Ding
Engineering Systems and Services
Delft University of Technology
Delft, Netherlands
aaron.ding@tudelft.nl

***Abstract*—Machine learning systems (MLSys) are emerging in the Internet of Things (IoT) to provision edge intelligence, which is paving our way towards the vision of ubiquitous intelligence. However, despite the maturity of machine learning systems and the IoT, we are facing severe challenges when integrating MLSys and IoT in practical context. For instance, many machine learning systems have been developed for large-scale production (e.g., cloud environments), but IoT introduces additional demands due to heterogeneous and resource-constrained devices and decentralized operation environment. To shed light on this convergence of MLSys and IoT, this paper analyzes the trade-offs by covering the latest developments (up to 2020) on scaling and distributing ML across cloud, edge, and IoT devices. We position machine learning systems as a component of the IoT, and edge intelligence as a socio-technical system. On the challenges of designing trustworthy edge intelligence, we advocate a holistic design approach that takes multi-stakeholder concerns, design requirements and trade-offs into consideration, and highlight the future research opportunities in edge intelligence.**

***Index Terms*—edge intelligence, machine learning systems, Internet of Things, trade-offs, trustworthiness, smart services**

I. Introduction

Machine learning systems are omnipresent and tireless silent helpers that bring order to our busy modern life: they guide us through traffic, classify and predict diseases in humans and plants, and are our eyes and ears in situations where we cannot see and hear. Their underlying machinery are machine learning algorithms that fit complex functions over data to discover patterns and correlations which can be exploited to discover trends and relationships, and for making predictions [1]. Many machine learning algorithms can be scaled to very large datasets and improve with more data. This has made them extremely successful in analysing the large volumes of data produced by digital, online services and applications. Deep neural networks in particular have produced state of the art results for many perception based tasks and are now widely used to process image, video, speech, audio and sequential data [2]. Machine learning is a promising technique when a system or process is not well understood, or too complex and difficult to model explicitly, but data that can surface insights about it has been collected [3]. Equally, if applications are dynamic and evolve over time, machine learning systems can use new data to discover patterns and update their predictions, thus adapting with the application.

The Internet of Things (IoT) [4] has matured from a vision of digitally connected devices to one of smart services [5] and ubiquitous intelligence. For example, a security camera that streams video footage to a remote server is no longer sufficient. Instead, the camera is expected to provide a smart service, such as counting people, or detecting an intruder, thus becoming an intelligent system rather than a mere device connected to the Internet. An intelligent system in the IoT distinguishes itself by having data processing capabilities [6], meaning that raw sensor data can be transformed to information and knowledge. This kind of abstraction allows humans to infer actionable insights about the system, which can be used to create services that add value to society [7].

Historically, human cognition has been needed to abstract information and knowledge from data. However, with the success of machine learning algorithms, new types of technology-driven intelligent systems are emerging that can deliver smart services with reduced human intervention. Machine learning systems are widely investigated to process sensor data and manage system performance and operation in the IoT [8]. They can be viewed from two perspectives: machine learning systems *for* the IoT support system management and organisation. These systems are designed in service of the IoT and use machine learning to improve overall system aspects like security [9], network traffic profiling and IoT device identification [10]. We focus on a second perspective in which machine learning systems are viewed as technical *components of* the IoT that perform advanced data processing tasks like activity recognition, object identification or keyword detection, in service of the greater application objective. We refer to this perspective as edge intelligence.

This paper motivates for an interdisciplinary approach that considers multi-stakeholder concerns, design requirements and trade-offs to develop trustworthy edge intelligence for smart services. In classical machine learning these are not considered, as reliable, abundant, scalable and almost-free communication networks and computing power under control and ownership of a single entity are assumed. In Section 2 we present an overview of machine learning systems and current concerns arising due to training, data, inference and operations. Section 3 highlights additional challenges that the IoT imposes on machine learning systems. In Section 4 we consider

edge intelligence trade-offs from a socio-technical and multi-stakeholder perspective. Section 5 presents an outlook for trustworthy edge intelligence, where we take the concerns arising both in machine learning systems and the IoT into consideration. We then highlight opportunities for trustworthy edge intelligence and finally conclude in Section 6.

II. Overview of Machine Learning Systems

Machine learning algorithms learn models from data by approximating useful functions that transform input variables, or features, to an output. This is called model training. Trained models are used to calculate an output value for a new input value, which is called inference [1]. When ground truth values, or labels, of the output values are available and used for training a model, the process is called supervised learning. A typical machine learning workflow involves data processing, model training and validation, and inference steps. Data processing serves two purposes. Firstly, input data is cleaned by removing outliers, missing values and errors. Secondly, data is transformed, for example by filtering the features that are used in the learning process. Machine learning systems must facilitate the ongoing deployment of machine learning workflows, which requires that they take operational aspects into consideration. This section highlights data, training, inference and operational concerns that currently challenge machine learning systems. Rather than being exhaustive, we aim to raise important considerations that are bound to impact supervised machine learning systems used in the IoT to provision smart services.

A. Model Training Concerns

Supervised model training is an iterative optimization problem that aims to find generalizable patterns in data. In statistical learning, the goal of model training is to find the candidate model from limited training data, that has the best predictive performance on new data [11]. Practically, the training process determines the values of those model parameters that minimize the error between a predicted value and its corresponding real value in the training data. In addition to parameter coefficients, a model can also have hyperparameters that control its complexity. To find the best model, a range of different model types, parameters and hyperparameters must be explored so that the best model can be selected. However, exploring each of these choices requires computational power, time and energy, resulting in trade-offs between predictive performance and resource consumption. State-of-the-art machine learning models, in particular deep neural networks, can have millions of parameters. Training them takes weeks or even months, and the computing and energy resources required are substantial. Two important approaches for improving the performance of machine learning systems are designing them together with specialized hardware and distributing model training for massive, parallel deployment across cloud servers [12].

In addition to the resources consumed during training, supervised machine learning requires labelled training data, which can be expensive and time consuming to collect [13]. When this is not possible, unsupervised learning which requires no ground truth labels, weak supervision with automated label generation [13], and approaches that reduce the amount of labels required [14] can be considered. Generally these present trade-offs against predictive performance. Due to the infrastructure requirements and cost of model training and data labelling, many applications download pre-trained models from online repositories, which can sometimes be used off-the-shelf, or otherwise adapted to new domains or datasets with transfer learning [15]. While pre-trained models speed up the development of new applications, they present significant security risks [16].

Classical machine learning algorithms are developed under static, benign, closed-world assumptions: they assume that the world does not change, that the environment is good-natured [17], and that all categories to be predicted were known during training and contained in the training data [18]. Obviously this does not correspond with reality. For example, in medical image classification, it has been established that training data can contain unrecognised categories that are not in the labels but that affect predictive outcomes [19]. While adversarial machine learning [20] can be used to improve the robustness of models under the malicious attack of an adversary, and lifelong learning [21] provides methods for continuous learning by accumulating and maintaining knowledge which can be used to improve future learning, the conditions under which different paradigms can be combined, and what vulnerabilities this may result in, are not obvious. Despite the success of machine learning algorithms, many challenges thus remain to train models that generalize well and have good predictive accuracy while also being resource-efficient, robust, and adaptive in new, real-world environments.

B. Data Provenance Concerns

The quality of a machine learning model is strongly influenced by the quality and underlying distribution of the data that was used to train it [22]. Due to this central role of data in machine learning, common features of raw observational data, like missing values, data redundancy and noise, significantly impact the performance of the model that is trained. Noise, for example, obscures the data signal and can result from random or systematic errors in the observations, or from data that has been tampered with. Model performance can be degraded further by propagating data errors that were generated during data processing through the entire machine learning workflow. To extend a software metaphor, such data errors are to machine learning systems what bugs are to code [23]. Once deployed, data discovery and management are a particular challenge. Datasets are often taken from different sources. As projects grow, so do dependencies between datasets. Over time the training data becomes increasingly complex to track and version [24]. Data dependencies and feedback loops are often hidden and can have unexpected effects that make machine learning systems brittle and error diagnosis expensive [25]. Machine learning systems are also vulnerable to attacks that exploit their dependency on data by polluting training data

(poisoning) or modifying input data before inference (evasion) [16].

C. Inference Concerns

Trained models infer output values for new data inputs to inform decisions or take actions on an ongoing basis. A trained model is a reusable asset that will make thousands or even millions of predictions before it is retrained. Unlike model training which happens in the background of an application, inference usually serves users directly and consequently needs to be efficient, reliable and interpretable. Even though the resource requirements for a single prediction are negligible in comparison to those of training a model, the scale at which inference happens requires efficient and optimized processes with high throughput, low latency and graceful performance degradation [26]. Traditionally, more attention has been devoted to optimizing the training process rather than inference. Recent releases of popular machine learning platforms like TensorFlow and MXnet now offer libraries for model optimization, but efficiency alone is not enough. When machine learning systems make decisions and act on our behalf, inference must also be reliable [27]. Current machine learning systems do not offer predictable throughput, latency and accuracy. Methods that guarantee model outputs and offer reliable uncertainty estimates are needed to provide inference with quality assurance [28]. Additionally, interpretable inference, which can be likened to the ability of humans to understand how a model works, is necessary for trusted, fair and ethical decision-making based on predictions [29].

D. Machine Learning Operations (ML Ops) Concerns

The code responsible for model training and inference is only a small component of the greater system, which includes components for configuration, data collection, data verification, feature extraction, machine resource management, analysis, process management, serving infrastructure and monitoring. Even though machine learning systems are constructed from these different components, models are not modular in the way that software is [25]. Model parameters are learned iteratively, and as dependent on the data distribution as on the features used for training and the hyperparameters. Due to these dependencies, individual models are not extensible and multiple models interact in non-obvious ways. However, models evolve as data changes, methods improve or software dependencies change [30]. Ongoing deployment, customisation, reuse and tracking are thus continuous challenges. Machine learning systems require end-to-end software support that facilitates the development, testing, configuration, deployment, management and maintenance of all components that affect data provenance, model training and inference [31].

III. IoT Challenges for Machine Learning Systems

Machine learning systems in smart services are constrained by the nature and requirements of the IoT: distributed, physically-bounded and resource-limited, wireless-connected computing devices that must deliver dynamic and context-aware functionality over multi-layered, heterogeneous architectures [7]. For machine learning systems this is both an opportunity and a challenge. By learning from data, they are well suited to offer IoT applications functionality that enables them to adapt to specific locations, environmental or social situations and to evolve with them over time. It could even be argued that machine learning systems are a prerequisite for delivering smart services at scale, as explicitly and perpetually defining and programming the logic for the IoT and its interactions with the physical world and social systems is impossible. At present, however, machine learning systems assume homogeneous and context-independent cloud computing infrastructure with scalable data processing and storage, uninterrupted and unrestricted power supply, and low latency and high bandwidth networks. This stable and consistent environment does not exist for the billions of connected devices in the IoT, where data offloading to wireless networks, distributed, heterogeneous computing infrastructure and resource-constrained devices with physical hardware limitations present trade-offs against each other and the performance of algorithms. To achieve scale, components in the IoT must also be reusable.

A. Offloading to Wireless Communication Networks

Wireless communication networks like Bluetooth and WiFi connect devices either as a local network, or they connect individual devices and local networks to the Internet [32]. Many IoT applications rely on wireless connections to offload data collected by devices to the cloud, where it can be cleaned and fused with other datasets, machine learning models can be trained, inference can be done and the data is stored for future use. Offloading gives access to greater computing power and storage, but poses privacy and performance concerns. Wireless communication links have a fixed throughput capacity and range [32], are lossy and noisy [33], and expose new attack surfaces [34]. Network interruptions are bound to affect IoT applications. At worst, machine learning systems must consider the risk of completely loosing connectivity during training or inference, making fault tolerance a necessary consideration [28]. At best, wireless connections introduce latency, variability, uncertainty and costs to machine learning systems, which historically have abstracted away their iterative communication requirements. Offloading thus weighs against privacy and real-time inference requirements, and constrains the frequency, size and data distribution of training updates of machine learning systems. While the data path, timing and transfer volumes can be optimized through routing schemes, scheduling and data compression to minimize bottlenecks and communication costs [35], this can reduce predictive accuracy and may be limited by the power supply and computing capabilities of devices [32].

B. Distribution Across Heterogeneous Devices

IoT endpoints (e.g. servers, sensors or mobile phones) that are located at the periphery of the Internet are called the edge.

Edge computing extends the computing power of the cloud to the endpoints [36], thus creating a geographically distributed network of processors for model training and inference. The edge varies in computing capabilities and connectivity from sensing and actuator devices that observe and control the environment at the lowest level, to gateways and cloud servers. Data processing, model training and inference on the edge can be device, gateway or cloud-centric [32]. Device-centric approaches reduce offloading challenges, but processing is limited by the computing capabilities and power supply of devices. Gateway-centric computation requires wireless communication, and introduces associated variability and uncertainty. Cloud-centric approaches offer unlimited storage and data processing capabilities, but come with copious communication overheads. Edge servers present an intermediate solution that offers stable power supply and processing closer to the points of data collection, while reducing the data transfer requirements that would be required by the cloud. A simple heuristic is that the availability of data processing, memory, storage and communication overheads all rise with increasing distance from devices. Increasing the former is desirable, while increasing communication overheads is not. The key challenge of distributing machine learning systems in the IoT is to decide whether, when and how to offload computations; that is, to find the optimal balance between local processing and computation offloading given unpredictable networks, and constrained and diverse devices and servers.

C. Resource Limited Devices

Battery-powered IoT devices have limited memory, processing and power supply and the resources that are available are shared between data collection, data processing (e.g. error detection, compression and encryption) and communication tasks [32]. Despite these limitations, on-device machine learning aims to do inference, partial model training and retraining locally on the device to remove the constraints associated with wireless communication. To make machine learning tasks in such resource constrained settings feasible, energy efficiency is of the essence [37]. The key requirements for this are to reduce the model size, the energy consumption and the processing requirements of model training and inference, while providing comparable predictive accuracy to what can be achieved on the cloud [38]. For mobile devices, federated learning [39] has become the standard for distributing model training. This approach reduces privacy concerns and data transfer volumes by processing sensitive data on devices and only performing global parameter aggregation in the cloud. Extensions to federated learning add differential privacy [40] to provide privacy guarantees. In federated learning systems for resource-constrained IoT networks, data transfer volumes, model training time and the temperature of devices present trade-offs [41].

Classical deep learning models can be several gigabits large. Small models are necessary for on-device inference for two reasons: on-device storage is low, and inference with larger models requires more computations, which consume more energy. To reduce the model size, quantization and pruning are used for model compression [42]. Quantization, which reduces the floating point precision of parameters and gradients, can be rule-based [43] or automated [44], with mixed bitwidths or optimized single bitwidth [45]. On the extreme end, binarized neural networks are quantized to 1, 2 or 3 bits [46] and provide superior efficiency, but at the cost of predictive accuracy. Mappings of binary neural networks to look-up tables on Field Programmable Gateway Arrays are able to reduce the energy consumption and latency even more [47]. Model pruning eliminates insignificant parameters from neural networks to reduce their size. Despite its popularity, advances in and the impact of model pruning are difficult to evaluate, as the field lacks standardized performance benchmarks [48]. A rising trend for on-device deep learning is the co-design of model and hardware architectures [49], and the exploration of a large search space of possible architectures with Neural Architecture Search [50].

D. Component Reusability

Reusability is an important design consideration in the IoT and a necessity for deploying smart services at massive scale [7]. This means that IoT components must be discoverable and useable by third parties to deploy new services. Components that lend themselves to reuse are hardware, data, models and the execution environments. For model training, raw data, features, sensing and processing devices can be shared. Similarly, for inference the sensors and processors, observational and transformed data streams, and models can be shared. Shared devices reduce the cost of hardware acquisition and system life-cycle cost (e.g. maintenance activities), which is an advantage. However, shared components bring their own challenges. Shared hardware and models challenge machine learning systems to consider hardware heterogeneity and utilization, workload allocation and prioritization, process scheduling and isolation, resource management and security. When many devices operate in close proximity, interference can affect data transmission, leading to increased energy consumption of devices, reduced service quality and communication delays. Shared data additionally poses questions of anonymity and control, governance and persistence: for example, who grants access to your phone's geolocation data to track your digital footsteps through the city? Do those that see your trail know it's you? And are you able to wipe your trace when you want to?

Sensing devices and smart services can be mapped in one-to-one, one-to-many, many-to-one and many-to-many configurations [32]. Training and inference workloads can be mapped to processing devices in a similar fashion. Collaborative inference with data inputs from multiple sensing devices, and multi-tenant processing which allocates and schedules multiple workloads over one or more resources, are the logical extension of pervasive sensing and edge intelligence to ubiquitous intelligence. Distributed machine learning operations for edge intelligence are bound to be complex and complicated. The heterogeneous and geographically dispersed IoT will amplify

the operational challenges already observed in the cloud. Moreover, sharing presupposes the involvement of multiple stakeholders, which inherently implies that ownership, governance, accountability and trust matter.

IV. Multi-Stakeholder Trade-Offs

The IoT is not only a complex collection of technologies, but a socio-technical system in a multi-stakeholder environment [51] with networks of independent actors consisting of users, data generators, network providers, data processors, application service providers and many others. With so many players involved, data and device use, management, maintenance and ownership are heterogeneous and can change. This multi-stakeholder environment gives rise to conflicting requirements and priorities between actors that must be considered when designing edge intelligence for smart services.

A. Design Aspects and Stakeholder Concerns in the IoT

Engineered systems are designed to deliver reliable, predictable and robust performance within acceptable bounds of confidence, in an unpredictable world. For example, boarding a plane when a thunderstorm is brewing, you have confidence that you will arrive at your destination because you have a justified belief that the plane was carefully designed, that it is operated by a well trained pilot and that the air traffic control system abides by internationally regulated standards of excellence. In its vision of smart services and ubiquitous intelligence, the IoT[1] serves as subsystem to larger, yet again socio-technical, engineered systems. Its hybrid cyber-physical nature however means that actions in the cyber realm carry consequences in the physical environment and can influence our experience of the world, like getting cold when a heating system is deactivated. This imposes more stringent requirements on its design than what would be the case for purely physical or solely cyber systems.

Specifications for the IoT are captured in standards (e.g. see references listed in [51]). A useful approach for identifying system requirements is through *concerns* that are of interest to one or more stakeholders [51]. Table I lists concerns, grouped into *aspects* based on common attributes, that have been developed to provide a comprehensive framework for the design of hybrid cyber and physical systems, like the IoT. Concerns are related and composable. For example, in considering the uncertainty concern, the latency imposed by specifying and managing uncertainty must also be considered. Typically concerns present trade-offs and stakeholders are likely to prioritize them differently. Requirements can be used to express system properties that address relevant concerns.

TABLE I
Aspects and concerns of IoT/CPS [51]

Aspects	Concerns
functional	actuation, communication, controllability, functionality, manageability, monitoriablity, performance, physical, physical context, sensing, states, uncertainty
business	enterprise, cost, environment, policy, quality, regulatory, time to market, utility
human	human factors, usability
trustworthiness	privacy, reliability, resilience, safety, security
timing	logical time, synchronization, time awareness, time-interval and latency
data	data semantics, identity, operations on data, relationship between data, data velocity, data volume
boundaries	behavioural, networkability, responsibility
composition	adaptability, complexity, constructivity, discoverability
lifecycle	deployability, disposability, engineerability, maintainability, operability, procurability, producibility

B. Implications for the Design of Edge Intelligence

Edge intelligence integrates machine learning systems into the cyber system of the IoT. Unlike the low risk analytical settings in which statistical machine learning has been developed, this can have real-world, potentially harmful or even life-threatening repercussions if the system malfunctions or fails. As a component of the IoT, it is thus necessary that machine learning systems for edge intelligence conform to the requirements of the IoT. And as with other software systems, specifying the target system behaviour during a requirements analysis process is essential. Machine learning systems in the wearables domain already incorporate explicit requirements analysis processes to specify system requirements upfront [53]. This is not the norm in other domains, and the opportunity exists to develop approaches for navigating conflicting design concerns and requirements trade-offs. These will need to consider the multi-layered and complex component technologies for edge intelligence, the limitations that they present individually and collectively, and the design choices that satisfy the prioritized requirements of stakeholders.

Opportunity: *Frameworks and processes are needed to elicit stakeholder requirements, navigate conflicting design concerns and prioritize trade-offs to make informed design choices for edge intelligence.*

True to its statistical heritage, the (implicit) design of machine learning systems in the IoT focuses primarily on feature engineering, algorithm selection, parameter optimization and architecture design, with the goal of optimizing predictive performance. From an IoT perspective, this addresses the performance concern of the functional aspect, but falls short on measuring and optimizing for other concerns. On-device machine learning (see Section III-C) already broadens concerns to account for physical contexts with resource limitations. Likewise, wireless offloading raises uncertainty, privacy, security, latency, data velocity and volume concerns, while distribution and device heterogeneity introduce con-

[1]The definitions of the IoT and cyber physical systems (CPS) have been converging over time [52]. We take a unified perspective of the two fields and refer to them collectively as IoT, to retain focus on machine learning systems.

trollability and synchronization concerns. Within a service paradigm, quality plays an important role, as it is viewed as a discriminating factor by which users choose services [7]. Providing ways for estimating uncertainty and for measuring, controlling and guaranteeing quality of service thus carry particular significance for smart services.

Opportunity: *Metrics and benchmarks beyond predictive performance are needed so that machine learning systems for edge intelligence can be specified, designed and evaluated.*

Issues of fairness, accountability and transparency are endemic to machine learning systems [54], where the data quality and distribution is integral to the model that is learnt. Models learned from data have the unfortunate drawback that they propagate the biases of the data collection process. Moreover, some machine learning algorithms, like deep neural networks, are considered to be "black box" algorithms, meaning that the inner workings of the algorithm according to which predictions are made are poorly understood and not controllable by humans. At present, IoT concerns do not consider concerns such as fairness, transparency, explainability and interpretability, that arise due to the data-centric nature of machine learning systems. They need to be accounted for to avoid becoming a blind spot in the design of edge intelligence.

Opportunity: *To be relevant to edge intelligence, concerns and aspects of the IoT need to be expanded to incorporate well known challenges due to the data-centric nature of machine learning systems.*

V. Outlook: Trustworthy Edge Intelligence

Trust-in-technology research extends trust beyond social systems to non-human, artificial entities. Technologies vary in their perceived "humanness", and users trust technologies differently based on this [55]. If the perceived humanness of a technology is high, then human-like trust constructs such as benevolence, integrity and ability, are good measures of trust. Congruently, if the perceived humanness is low, then system-like trust constructs like helpfulness, reliability and functionality are more appropriate measures. While related, trust and trustworthiness represent different concepts [56]. Trust is a psychological state that indicates whether a trustor is willing to take risks for a trustee in the absence of monitoring or external control. Trustworthiness is a necessary condition for choosing to trust someone and focuses on the characteristics of a trustee. Trustworthiness concerns are essential considerations in both AI and the IoT, but they are approached from different perspectives in the two fields.

A. Trustworthiness in AI

Technologies that create the perception of social presence of other humans, that facilitate social behaviour (e.g. engaging in dialogue or receiving affection), and that enable interactions with other people are perceived to be more human-like [55]. Artificial intelligence technologies, which encompass machine learning systems, are thus human-like by definition and design. Heightening public mistrust has lead governments and organisations to rapidly develop AI frameworks to specify principles for trustworthy AI. The core themes that emerge from prominent frameworks are: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values [57]. These trust constructs resonate with the perspective that AI technologies are perceived to be human-like. While the frameworks lay the theoretical ground work, to be useful trustworthy AI needs to develop measurable trustworthiness concerns that can lead to practical and enforceable specifications.

Opportunity: *Trustworthiness concerns of machine learning systems need to be standardized and operationalized so that they can be incorporated in specifications and evaluated objectively in applications.*

B. Trustworthiness in the IoT

In contrast, trustworthiness concerns in the IoT are agreed on across the industry, captured in standards, and formally defined as safety, security, privacy, reliability and resilience [51]. The concerns serve to assure that systems behave as expected under various operating conditions. They support the view that the IoT is perceived to be less human-like, and more system-like. Other properties such as controllability, manageability, functionality, performance and uncertainty are considered as functionality concerns, rather than trustworthiness concerns.

C. Opportunities for Trustworthy Edge Intelligence

Neither trustworthy AI, nor trustworthiness concerns in the IoT address the full spectrum of trustworthiness concerns that arise in edge intelligence. For example, a machine learning system may fail to make correct predictions under open world assumptions, which can include new categories, unseen examples, black swan events and foreign attack models. To be able to perform fault diagnosis in such scenarios, explainability is a necessary requirement. Or consider a smart camera installed in a new context where the population does not resemble the people that were represented in the training data of the deployed model. The machine learning system may fail to recognize members of that population and the trust constructs of fairness and non-discrimination will directly impact the functionality of the application. On the other hand, intermittent and unreliable data transfer over wireless channels can result in missing values that limit inference quality and affect system level predictive performance. A voice assistant that alerts emergency response when you cry for help, will need to perform reliably even in those settings. Trustworthy edge intelligence thus requires that trust constructs for machine learning systems and trustworthiness concerns arising in the IoT are considered together. As with other design requirements, trustworthiness concerns will be composable and pose trade-offs against each other and against other stakeholder concerns. There is thus a need to:

- analyze the trustworthiness concerns that arise in machine learning systems for edge intelligence and smart services
- explore the overlap and trade-offs of trustworthiness concerns between machine learning and IoT systems

- characterize the interactions and trade-offs between trustworthiness concerns and other stakeholder concerns
- expand research into trustworthy machine learning to also address the diverse spectrum of challenges and trade-offs that arise in edge intelligence

VI. Concluding Remarks

Ever-growing, densely populated urban centers need to monitor, track, care for and nurture their social, natural and artificial systems. Smart services, informed by ubiquitous intelligence, are viewed as a way of doing this. Machine learning systems can enable smart services by provisioning the IoT with edge intelligence, giving rise to ubiquitous intelligence. This paper presents challenges and trade-offs that arise when designing trustworthy edge intelligence for smart services. Despite the maturity of machine learning systems and the IoT, combining the two technologies presents new concerns for edge intelligence. One the one hand, many machine learning systems have been deployed in large-scale production environments, and the model training, data provenance, inference and ongoing operational challenges are known. These challenges prevail when deploying machine learning systems in the IoT, but are not considered in existing IoT design frameworks. On the other hand, additional challenges arise due to communication offloading, distributed, heterogeneous and resource-constrained devices, and the need to share and reuse components in the IoT. These challenges are not addressed by classical machine learning, or large scale, cloud-based machine learning systems.

We position machine learning systems as a component of the IoT, and edge intelligence as a socio-technical system. We motivate that multi-stakeholder concerns, design requirements and technology trade-offs should be taken into consideration when developing edge intelligence, and highlight opportunities that exist to facilitate this. With an outlook on trustworthiness, we demonstrate that an interdisciplinary perspective is essential, as trust constructs are considered differently in machine learning systems and the IoT. By combining perspectives, and taking multi-stakeholder concerns, design requirements and trade-offs into considerations, it is possible to perceive of a future where holistic, trustworthy edge intelligence and smart services are possible.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009. [Online]. Available: https://web.stanford.edu/ hastie/ElemStatLearn/

[2] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] T. M. Mitchell, *The discipline of machine learning*. Carnegie Mellon University, School of Computer Science, Machine Learning ..., 2006, vol. 9.

[4] R. Minerva, A. Biru, and D. Rotondi, "Towards a Definition of the Internet of Things (IoT)," *IEEE Internet Initiative*, pp. 1–86, 2015.

[5] D. Georgakopoulos and P. Prakash Jayaraman, "Internet of things: from internet scale sensing to smart services," *Computing*, vol. 98, pp. 1041–1058, 2016. [Online]. Available: https://doi.org/10.1007/s00607-016-0510-0

[6] J. E. Ibarra-Esquer, F. F. González-Navarro, B. L. Flores-Rios, L. Burtseva, and M. A. Astorga-Vargas, "Tracking the evolution of the internet of things concept across different application domains," *Sensors (Switzerland)*, vol. 17, no. 6, pp. 1–24, 2017.

[7] A. Bouguettaya, B. Medjahed, M. Ouzzani, F. Casati, X. Liu, H. Wang, D. Georgakopoulos, L. Chen, S. Nepal, Z. Malik, A. Erradi, M. Singh, Y. Wang, B. Blake, S. Dustdar, F. Leymann, M. Papazoglou, M. Huhns, Q. Sheng, H. Dong, Q. Yu, A. G. Neiat, S. Mistry, and B. Benatallah, "A service computing manifesto: the next 10 years," *Communications of the ACM*, vol. 60, no. 4, pp. 64–72, 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/2983528

[8] F. Samie, L. Bauer, and J. Henkel, "From cloud down to things: An overview of machine learning in internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4921–4934, 2019.

[9] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 3, pp. 1646–1685, 7 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9072101

[10] L. Cui, S. Yang, F. Chen, Z. Ming, N. Lu, and J. Qin, "A survey on application of machine learning for Internet of Things," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 8, pp. 1399–1417, 2018. [Online]. Available: http://dx.doi.org/10.1007/s13042-018-0834-5

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Scholkopf, Eds. Springer, 2006.

[12] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM Comput. Surv.*, vol. 53, no. 2, Mar. 2020. [Online]. Available: https://doi.org/10.1145/3377454

[13] A. Ratner, C. De Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: Creating large training sets, quickly," *Advances in Neural Information Processing Systems*, no. Nips, pp. 3574–3582, 2016.

[14] C. Renggli, B. Karlaš, B. Ding, F. Liu, K. Schawinski, W. Wu, C. Zhang, and S. Grünewälder, "Continous Integration of Machine Learning Models with ease.ml/ci: Towards a Rigorous yet Practical Treatment," in *Proceedings of the 2nd SysML Conference*, Palo Alto, US, 2019. [Online]. Available: http://arxiv.org/abs/1903.00278

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5288526

[16] Y. Ji, X. Zhang, S. Ji, X. Luo, and T. Wang, "Model-reuse attacks on deep learning systems," *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 349–363, 2018.

[17] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Communications of the ACM*, vol. 61, no. 7, pp. 56–66, 7 2018. [Online]. Available: https://dl.acm.org/doi/fullHtml/10.1145/3134599

[18] G. Fei, S. Wang, and B. Liu, "Learning cumulatively to become more knowledgeable," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu. New York, NY, USA: Association for Computing Machinery, 8 2016, pp. 1565–1574. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939835

[19] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré, "Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging," in *Proceedings of the ACM Conference on Health, Inference, and Learning*. New York, NY, USA: ACM, 2020. [Online]. Available: https://doi.org/10.1145/3368555.3384468

[20] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial Machine Learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, 2011, pp. 43–58. [Online]. Available: https://doi.org/10.1145/2046684.2046692

[21] Z. Chen and B. Liu, "Lifelong Machine Learning," in *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 1st ed., R. J. Brachman, W. W. Cohen, and P. Stone, Eds. Morgan & Claypool, 2016.

[22] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[23] E. Breck, N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data Validation for Machine Learning," in *Proceedings of the 2nd SysML Conference*, 2019, pp. 1–14.

[24] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software Engineering for Machine

Learning: A Case Study," *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP 2019*, pp. 291–300, 2019.

[25] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden Technical Debt in Machine Learning Systems," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2015, pp. 2503–2511.

[26] Y. Lee, A. S. Politecnico, D. Milano, B.-G. Chun, M. Domenico, S. Politecnico, and M. W. Microsoft, "PRETZEL: Opening the Black Box of Machine Learning Prediction Serving Systems," in *Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation*, 2018. [Online]. Available: https://www.usenix.org/conference/osdi18/presentation/lee

[27] I. Stoica, D. Song, R. A. Popa, D. Patterson, M. W. Mahoney, R. Katz, A. D. Joseph, M. Jordan, J. M. Hellerstein, J. E. Gonzalez, K. Goldberg, A. Ghodsi, D. Culler, and P. Abbeel, "A Berkeley View of Systems Challenges for AI," 2017. [Online]. Available: http://arxiv.org/abs/1712.05855

[28] T. Abdelzaher, Y. Hao, K. Jayarajah, A. Misra, P. E. R. Skarin, S. Yao, and D. Weerakoon, "Five Challenges in Cloud-enabled Intelligence and Control," *ACM Transactions on Internet Technology*, vol. 20, no. 1, pp. 1–19, 2020.

[29] Z. C. Lipton, "The of Model The Interpretability," pp. 1–28, 2018.

[30] S. Schelter, F. Biessmann, T. Januschowski, D. Salinas, S. Seufert, and G. Szarvas, "On Challenges in Machine Learning Model Management," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pp. 5–13, 2018. [Online]. Available: http://sites.computer.org/debull/A18dec/p5.pdf

[31] A. Ratner, D. Alistarh, G. Alonso, D. G. Andersen, P. Bailis, S. Bird, N. Carlini, B. Catanzaro, J. Chayes, E. Chung, B. Dally, J. Dean, I. S. Dhillon, A. Dimakis, P. Dubey, C. Elkan, G. Fursin, G. R. Ganger, L. Getoor, P. B. Gibbons, G. A. Gibson, J. E. Gonzalez, J. Gottschlich, S. Han, K. Hazelwood, F. Huang, M. Jaggi, K. Jamieson, M. I. Jordan, G. Joshi, R. Khalaf, J. Knight, J. Konečný, T. Kraska, A. Kumar, A. Kyrillidis, A. Lakshmiratan, J. Li, S. Madden, H. B. McMahan, E. Meijer, I. Mitliagkas, R. Monga, D. Murray, K. Olukotun, D. Papailiopoulos, G. Pekhimenko, T. Rekatsinas, A. Rostamizadeh, C. Ré, C. De Sa, H. Sedghi, S. Sen, V. Smith, A. Smola, D. Song, E. Sparks, I. Stoica, V. Sze, M. Udell, J. Vanschoren, S. Venkataraman, R. Vinayak, M. Weimer, A. G. Wilson, E. Xing, M. Zaharia, C. Zhang, and A. Talwalkar, "MLSys: The New Frontier of Machine Learning Systems," pp. 1–4, 2019. [Online]. Available: http://arxiv.org/abs/1904.03257

[32] F. Samie, L. Bauer, and J. Henkel, "IoT Technologies for Embedded Computing : A Survey," in *CODES/ISSS '16*, Pittsburgh, USA, 2016.

[33] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications Ala," *IEEE COMMUNICATION SURVEYS & TUTORIALS*, vol. 17, no. 4, p. 2347, 2015.

[34] S. Sen, J. Koo, and S. Bagchi, "Trifecta: Security, energy efficiency, and communication capacity comparison for wireless iot devices," *IEEE Internet Computing*, vol. 22, no. 1, pp. 74–81, 2018.

[35] H. Luo, Y. Liu, and S. K. Das, "Routing Correlated Data with Fusion Cost in Wireless Sensor Networks," *IEEE Transactions on Mobile Computing*, vol. 5, no. 11, pp. 1620–1632, 2006.

[36] I. Sittón-Candanedo, R. S. Alonso, J. M. Corchado, S. Rodríguez-González, and R. Casado-Vara, "A review of edge computing reference architectures and a new global edge proposal," *Future Generation Computer Systems*, vol. 99, no. 2019, pp. 278–294, 2019. [Online]. Available: https://doi.org/10.1016/j.future.2019.04.016

[37] E. Farella, M. Rusci, B. Milosevic, and A. L. Murphy, "Technologies for a thing-centric internet of things," *Proceedings - 2017 IEEE 5th International Conference on Future Internet of Things and Cloud, FiCloud 2017*, vol. 2017-Janua, pp. 77–84, 2017.

[38] C. R. Banbury, V. J. Reddi, M. Lam, W. Fu, A. Fazel, J. Holleman, X. Huang, R. Hurtado, D. Kanter, A. Lokhmotov, D. Patterson, D. Pau, J.-s. Seo, J. Sieracki, U. Thakker, M. Verhelst, and P. Yadav, "Benchmarking TinyML Systems: Challenges and Direction," 2020. [Online]. Available: http://arxiv.org/abs/2003.04821

[39] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, vol. 54, 2017.

[40] H. B. McMahan, G. Andrew, U. Erlingsson, S. Chien, I. Mironov, N. Papernot, and P. Kairouz, "A General Approach to Adding Differential Privacy to Iterative Training Procedures," 2018. [Online]. Available: http://arxiv.org/abs/1812.06210

[41] A. Feraudo, P. Yadav, V. Safronov, D. A. Popescu, R. Mortier, S. Wang, P. Bellavista, and J. Crowcroft, "CoLearn: Enabling Federated Learning in MUD-compliant IoT Edge Networks," in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking (EdgeSys '20)*, vol. 20, 2020. [Online]. Available: https://doi.org/10.1145/3378679.3394528

[42] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," 2016. [Online]. Available: http://arxiv.org/abs/1602.07360

[43] M. Rusci, A. Capotondi, and L. Benini, "Memory-Driven Mixed Low Precision Quantization for Enabling Deep Network Inference on Microcontrollers," in *Proceedings of the 3rd MLSys Conference*, Austin, Texas, 2020.

[44] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "HAQ: Hardware-aware automated quantization with mixed precision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 8604–8612, 2019.

[45] T.-W. Chin, P. I.-J. Chuang, V. Chandra, and D. Marculescu, "One Weight Bitwidth to Rule Them All," 2020. [Online]. Available: https://arxiv.org/abs/2008.09916

[46] J. Fromm, M. Cowan, M. Philipose, L. Ceze, and S. Patel, "Riptide: Fast end-to-end Binarized Neural Networks," in *Proceedings of the 3rd MLSys Conference*, Austin, Texas, 2020.

[47] S. Chidambaram, J. M. P. Langlois, and J. Pierre, "PoET-BiN: Power Efficient Tiny Binary Neurons," in *Proceedings of the 3rd MLSys Conference*, Austin, US, 2020.

[48] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag, "What is the State of Neural Network Pruning," in *Proceedings of the 3rd MLSys Conference*, Austin, US, 2020.

[49] D. Stamoulis, R. Ding, D. Wang, D. Lymberopoulos, B. Priyantha, J. Liu, and D. Marculescu, "Single-Path NAS: Designing Hardware-Efficient ConvNets in Less Than 4 Hours," in *ECML PKDD 2019: Machine Learning and Knowledge Discovery in Databases*, vol. 11907 LNAI. Springer, 9 2019, pp. 481–497. [Online]. Available: https://doi.org/10.1007/978-3-030-46147-8_29

[50] B. Zoph and Q. V. Le, "Neural Architecture Search with Reinforcement Leaning," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–16, 2017.

[51] CPS Public Working Group, "Framework for Cyber-Physical Systems : Volume 1 , Overview," National Institute of Standards and Technology, Tech. Rep., 2017.

[52] C. Greer, M. Burns, D. Wollman, and E. Griffor, "Cyber-Physical Systems and Internet of Things NIST Special Publication 1900-202," National Institue of Standards and Technology, Tech. Rep., 2019.

[53] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, and R. Jafari, "Enabling effective programming and flexible management of efficient body sensor network applications," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 1, pp. 115–133, 2013.

[54] H. Wallach, "Big data, machine learning, and the social sciences: Fairness, accountability, and transparency," 2014. [Online]. Available: https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d

[55] N. K. Lankton, D. Harrison Mcknight, and J. Tripp, "Technology, humanness, and trust: Rethinking trust in technology," *Journal of the Association for Information Systems*, vol. 16, no. 10, pp. 880–918, 2015.

[56] Y. J. Cho and J. W. Lee, "Perceived Trustworthiness of Supervisors, Employee Satisfaction, and Cooperation," *Public Management Review*, 2011. [Online]. Available: https://doi.org/10.1080/14719037.2011.589610

[57] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI," Berkman Klein Center for Internet & Society, Tech. Rep., 2020.