



# Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'

Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee and Jatinder Singh

Compliant & Accountable Systems Group, Dept. of Computer Science & Technology

University of Cambridge, UK

firstname(s).lastname@cst.cam.ac.uk

## ABSTRACT

AI is increasingly being offered 'as a service' (AIaaS). This entails service providers offering customers access to pre-built AI models and services, for tasks such as object recognition, text translation, text-to-voice conversion, and facial recognition, to name a few. The offerings enable customers to easily integrate a range of powerful AI-driven capabilities into their applications. Customers access these models through the provider's APIs, sending particular data to which models are applied, the results of which returned.

However, there are many situations in which the use of AI can be problematic. AIaaS services typically represent generic functionality, available 'at a click'. Providers may therefore, for reasons of reputation or responsibility, seek to ensure that the AIaaS services they offer are being used by customers for 'appropriate' purposes.

This paper introduces and explores the concept whereby AIaaS providers uncover situations of possible service misuse by their customers. Illustrated through topical examples, we consider the technical usage patterns that could signal situations warranting scrutiny, and raise some of the legal and technical challenges of monitoring for misuse. In all, by introducing this concept, we indicate a potential area for further inquiry from a range of perspectives.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Computer systems organization** → **Cloud computing**; • **Social and professional topics** → **Computing / technology policy**.

## KEYWORDS

artificial intelligence; machine learning; cloud computing; law; accountability; misuse; monitoring; audit; compliance; MLaaS; AIaaS

## ACM Reference Format:

Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee and Jatinder Singh. 2020. Monitoring Misuse for Accountable 'Artificial Intelligence as a Service'. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, February 7–8, 2020, New York, NY, USA. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3375627.3375873>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions.acm.org](https://permissions.acm.org).

AIES '20, February 7–8, 2020, New York, NY, USA

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7110-0/20/02...\$15.00

<https://doi.org/10.1145/3375627.3375873>

## 1 INTRODUCTION

Artificial Intelligence (AI) has recently seen a surge of interest. It is touted to influence many aspects of modern society, including in areas such as health, agriculture, finance, transport, manufacturing, retail, education, science, government and public services, to name but a few.

At the same time, AI is coming under increasing scrutiny [5, 30]. As the discussions of 'algorithmic accountability', 'AI regulation' and 'ethical AI' make clear, there are many situations in which the use of AI will be inappropriate, controversial and unlawful [16].

We increasingly see providers offering 'AI as a Service' (AIaaS). In essence, AIaaS aims at providing the ML (machine learning)-driven building blocks to support customer applications [26]. It includes service providers offering access to a range of pre-built models and related services, whereby customers can send to the service particular inputs and receive back the results of a ML process, e.g. predictions, classifications, etc. Offerings tend to be fairly generic<sup>1</sup>—examples including object detection, text to voice synthesis, facial recognition, and text translation (see §2.1)—and therefore can be attractive to a range of customers, and suitable for driving applications across a number of scenarios and sectors. For instance, the same object recognition service used by one customer for warehousing might be used by another to support video surveillance.

Interesting considerations are raised by AIaaS, given the services make available sophisticated AI capabilities, often as 'turnkey' (on demand, with a few clicks), to potentially anyone. That is, *there is much scope for AIaaS services to be used for controversial and problematic purposes*. As a provocative example, facial recognition is an AIaaS service offered by several providers, and intuitively has huge potential for misuse and abuse [10]. In line with this, we have seen providers setting out principles of use [21], being active in related public discourse [13], and implementing simple technical barriers to limit misuse (see §2). However even seemingly benign services also have the potential to drive controversial applications, e.g. generic object recognition services can assist military operations [38].

This paper provides an initial exploration into monitoring as a means for facilitating greater oversight and governance of the use of AIaaS. Providers appear to have an interest in such, be it for reasons of reputation, e.g. to avoid the backlash where an unsavoury application could be labelled as 'Powered by [OrgX]' or '[ProviderY] Inside™'; or more generally, to help ensure that customer usage accords with their terms of service, principles, or requirements. Our concept of monitoring service usage might also work to inform future legal or regulatory regimes, which may demand more accountability from those providing access to AI/ML services.

<sup>1</sup>Though tailoring is possible, and some offer consulting services for specific needs.

Specifically, we explore how AIaaS usage patterns could indicate situations warranting attention, and highlight challenges and opportunities for future work in the legal and technical spaces.

## 2 AIAAS OVERVIEW

AI is a broad term. In the context of AIaaS, it typically refers to the use of machine learning (ML) (though we use ‘AI’ and ‘AIaaS’ to be consistent with the terminology used by providers in this space). ML works to uncover patterns in data, to build and refine representative *models* of that data [28]. These models can be used to make classifications, predictions, and so forth.

There is significant interest in using ML to underpin a range of applications. However, undertaking ML, i.e. building models, can be challenging [41]. Organisations may have issues regarding access to data – given that models are built on data, one requires access to sufficient volumes of data to be representative of the problem space, and to enable model training and testing [8]. There are also issues of access to resources. Training models can be computationally expensive [8], so access to sufficient compute is a concern; as is ML expertise, which is said to be in short supply [19].

Recognising an opportunity, we increasingly see providers offering AI as a Service (AIaaS). The motivation for AIaaS is that many organisations will seek to leverage AI within their applications, but may lack the data, resources, capabilities, or time to undertake and maintain everything in-house. Indeed, this is similar to what drives the uptake of cloud infrastructure more generally. Unsurprisingly, we see that the prominent AIaaS providers are the tech giants, given their access to data, infrastructure and expertise.

Broadly there are two categories of AIaaS, those that (i) provide the infrastructure to support customers in undertaking their own machine learning; and those that (ii) provide access to a range of pre-built models and related services, whereby customers can send to the service particular inputs and receive back results, e.g. predictions, classifications, etc. Here we focus on the latter, using ‘AIaaS’ to refer to such (though some of the aspects we raise may also be relevant to the more model building-oriented ML services).

In this way, by offering access to models, AIaaS essentially provides ‘application building blocks’ which enables customers (tenants) to more easily integrate AI capabilities into the systems they build and run. One can expect AIaaS to grow in prominence. This is not only due to the increasing interest in AI, but also because customers benefit through lower barriers to entry: on-demand and pay-per-use reduces overheads in terms of engineering effort, cost and time-to-market, and where issues of maintenance and scalability are ‘outsourced’ to providers.

### 2.1 Types of services

AIaaS entails service providers offering their customers (tenants) access to models via APIs (Application Programming Interfaces – these facilitate the interaction between systems).

Typically, AIaaS offerings tend towards more generic tasks. We observe that several AIaaS providers [12, 20], group their offerings into four main categories:

**Language** services include text analytics (e.g. sentiment analysis), translation (e.g. automated text translation like Google Translate,

language detection, etc.), language understanding and knowledge base creation (e.g. from collections of questions and answers).

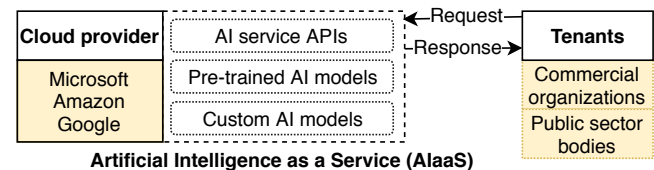
**Analytics** services comprise capabilities for the analysis of data for various purposes, such as product recommendations (e.g. to deliver personalised ads), anomaly detection (e.g. for detecting when data behaviour changes), knowledge inference (e.g. to make predictions or forecasts based on your data) and content moderation (e.g. to detect offensive or unwanted images).

**Speech** services primarily comprise features such as text-to-speech, speech-to-text, speaker recognition (i.e. identifying individuals based on patterns in their speech), and so forth.

**Vision** services enable the analysing of images and videos in order to find and identify objects, text, and labels, just to name a few. Relevant to our later discussion, there are also offerings relating to faces: Microsoft’s Azure Face [20] offers various APIs including face verification, face detection, emotion recognition, face identification, similar face search, celebrity recognition; Amazon Rekognition offers a similar range of face-oriented services [2].

### 2.2 Accessing AIaaS

To integrate AI into their applications, AIaaS customers setup and configure the services with possible customisations (which can entail model training). This is usually done through a specialised web interface supplied by the provider. Usage of the AIaaS service entails sending requests (e.g. ‘find all faces in this image’) to the provider’s API. The provider receives a request, and after conducting the required authentication and authorisation checks, processes the request given and returns the response (see Fig. 1).



**Figure 1: A simplified illustration of AIaaS in which cloud providers make AI technologies accessible to their customers (tenants).**

### 2.3 Transaction-oriented pricing

We explored the pricing model of three major providers, namely Amazon, Microsoft and Google, focusing on their vision-related AIaaS offerings. Despite subtle differences, the pricing structures are largely similar across the providers. In general, the API calls made can request one or more AI services (e.g., label detection, face recognition, object detection, etc.) to be applied to their input images. Azure’s vision pricing model [20] refers to each service call as a *transaction*. Several AIaaS service requests can be grouped into a single API call, but the tenant will be charged according to the total number of services requested. The Amazon Rekognition Image API pricing model for faces [2] charges per image analysed and ‘for each set of facial feature vectors you store’. Google charges per image [12], where ‘each feature applied to an image is a billable

unit', again meaning that the cost incurred per image will vary depending on the number of services applied to that image.

In short, all the major providers essentially bill in line with usage, relating to API calls (requests/responses). This is relevant as such billing methods naturally entail some monitoring, if only for accountancy purposes. As such, billing may provide a starting point for monitoring for misuse, as it implies providers already employ some means for monitoring transactions, and that some data is available regarding the transactions themselves.

### 3 POTENTIAL FOR MISUSE

The potential for the misuse of AI in general has been the subject of much attention. There are a number of examples of inappropriate uses of machine learning. One example is the infamous attempt to use deep learning to predict a person's sexual orientation from a photograph, which was widely criticised for many reasons, including that such applications would have much potential for misuse in terms of supporting persecution and hate crime [22]. In the facial recognition context, there has been much debate regarding the acceptability of such technologies, leading some governments to institute a ban on their use by the public sector [6]. The recent media coverage of police use of facial recognition on protesters in Hong Kong [39] has sparked fears of persecution. The installation of facial recognition systems in London raised privacy concerns [7].

Given the discussion around ethical AI, algorithmic accountability, and so forth, technology companies have joined efforts in self-regulation [29], and defining principles regarding AI. For instance, the Partnership on AI, founded by Amazon, Facebook, Google, DeepMind, Microsoft, and IBM—who incidentally are the prime AIaaS providers—lists six core tenets of AI: 1) safety-critical, 2) fair, transparent, and accountable, 3) labor and economy, 4) collaboration with people, 5) social and societal influences, and 6) AI and social good [29]. Many have their own principles of AI ethics; Google's states that AI should protect privacy and be socially beneficial, fair, safe, and accountable to people, while IBM's focuses on trust and transparency [29].

Despite this attention on the potential risks of AI, there has been considerably less focus on AI when offered as a service, and the potential for cloud-based AI services to be misused by tenants.

We argue that this warrants consideration. We have seen that seemingly innocuous technology can quickly be re-purposed as the building blocks for nefarious applications. 'Deepfakes', the use of AI to superimpose another person's face in a video, can be easily built using Google's open-sourced software with readily accessible tutorials, raising risks that it will be used to spread misinformation, threaten national security, or create pornography depicting non-consenting individuals [14].

AIaaS presents a challenge, given that it deliberately aims to make widely accessible sophisticated AI that is capable of underpinning a broad range of applications. This 'out-of-the-box' access to such functionality makes it likely that AIaaS will be used, maliciously or otherwise, by some customers for controversial purposes.

#### 3.1 Monitoring AIaaS

There are various reasons why providers might become active in monitoring and auditing how their AIaaS offerings are being used.

Monitoring and audit supports providers in taking a proactive role, by indicating situations that require attention and enabling the mitigation of negative outcomes.

A key driver is *reputation*. Given the potential concerns regarding the use of AI for particular purposes, a controversial application being branded as '*Powered by [Provider\_X]*' could lead to public backlash and undesirable attention. This is particularly the case for AIaaS, given that such services provide access to fairly generic models that could be used for a wide variety of purposes.

Providers also typically define in their terms of services particular responsibilities and obligations of use. Providers may therefore be encouraged to employ some monitoring as a means for indicating situations where breaches may be taking place. Moreover, given that much value lies within the models being accessed, providers may seek to prevent data being leaked or aspects of the model being 'stolen', e.g. through model inversion attacks [32].

Indeed, there is evidence of AIaaS providers beginning to consider such issues, though so far this has tended only to concern specific, more obviously 'risky' offerings – particularly facial recognition. For example, Google does not provide facial recognition as a service due to its potential for abuse [37], while Microsoft has been involved in the broader public discussion around such issues [31] and denied police access to its facial recognition technology [17]. In terms of technical measures, Microsoft limits the request rate to their Face API, while Amazon prevents more than 100 faces from being detected in single image [2], both of which may work to mitigate misuse. That said, more general frameworks regarding the appropriate uses of AIaaS, and how these would be monitored, so far have had little discussion.

Conversely, providers may be hesitant to become more involved in monitoring the use of their services. First, their customers (tenants) may not accept their activity being monitored, perhaps due to, for example, the nature or commercial sensitivity of their undertakings. Indeed, trust regarding cloud providers is a long-standing issue [27]. Further, monitoring could result in an increased liability exposure for the provider (see §5.1).

However, there is also an argument that AIaaS providers *should* play a greater role in monitoring their services. In addition to the points just mentioned, this is because providers (and indeed, their customers) profit from the offering of low-cost access to powerful ML services for use at scale. Moreover, while the misuse of cloud infrastructure services may be a general concern [15, 18], AIaaS warrants particular consideration, not only due to the potential for AI misuse, but also because such services mean the provider plays a more direct role in enabling application functionality. As such, one could envisage providers incentivised to take a more proactive role, be it through their own volition, given the demands for greater tech-accountability, or as a result of the evolving regulatory landscape. We discuss some legal aspects in greater depth in §5.1, though reiterate that the purpose of this paper is to introduce the concept and the potential of AIaaS monitoring and audit, in order to provide the groundwork for future discussion and research.

### 4 MONITORING FOR AIaaS MISUSE

There are several ways that providers could employ tighter governance regimes in AIaaS contexts. This could include, for instance,

**Table 1: Possible information sources to support the auditing of AlaaS usage.**

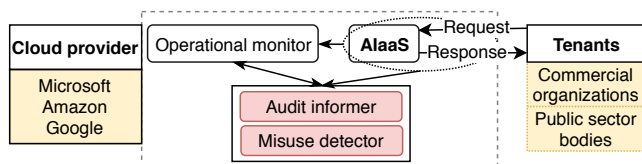
Category	Sub-category	Examples
Transaction raw data	Metadata	Start time, client IP address, size of input parameters, resource usage, ...
	Request	Input parameters, such as images, videos, audio files, ...
	Response	Information returned to the clients during a transaction (e.g., detected faces)
Request processing by-products		Generated information (e.g., face encoding) as part of the request processing
Derived data		Data structures or ML models obtained from the raw and by-product data

changing how services are offered, such as requiring initial consulting and application vetting rather than ‘turnkey’ service access (as some do), requiring specifications of use inline with well-defined and context-specific terms of service provisions, and so forth.

However, given the current and well-established model for cloud (and AlaaS) procurement—which is ‘turnkey’, on-demand, through automatic, tech-driven processes—we consider how providers might technically discover that their AlaaS services are being used inappropriately. Our focus is in line with the current way AlaaS operates, where providers are fairly ‘hands-off’ with regards to how their services are procured and integrated by customers.

In essence, our concept involves exploring how *monitoring and analysing customer usage of AlaaS services may help providers uncover situations warranting review or further investigation*. We emphasise that such an approach, which involves examining patterns of use, transactions, etc., will often only be indicative of possible misuse rather than absolute, yet remains useful in highlighting potential problems as they occur and in enabling responses. This offers much potential compared to the alternative of ‘running blind’.

Fig. 2 presents a high-level conceptual representation of a generic AlaaS misuse detection architecture. The concept illustrates a component called *Audit Informer*, which obtains and derives the relevant data to drive the analysis and detection of potential situations of concern. Some of this data might relate to that already collected by providers (via the *Operational Monitor*) in order to, for example, enable billing (which is transaction oriented, see §2.3), managing performance and other resources, etc. The *Misuse Detector* component involves analysing and processing the audit information as a means for identifying and alerting of potential AlaaS misuse.

**Figure 2: A conceptual AlaaS misuse detection architecture.**

#### 4.1 Audit information

Detecting AlaaS misuse requires information. There may be a range of ‘signals’ available to a cloud provider. Table 1 presents some categories of information that may be relevant, which include:

- *Transaction raw data*: Includes three main sub-categories, namely request metadata (e.g., timestamp, size, account credentials), request specifics (e.g., input parameters such as

images, audio clips, etc.), and response specifics (e.g., outputs from the model such as predictions or classifications). Note that some of this may already be collected by providers (Operation Monitor), e.g. to support billing.

- *Request processing by-products*: Includes data generated in serving the request, which can contain, for example, any input data pre-processing, internal processing pipelines, collation of results from multiple services, intermediary feature vectors, etc.
- *Derived data*: Concerns data processed or derived from transaction logs, request by products, etc. to assist detection and analysis. This could include, for example, creating structured log files, bloom filters, models and representations of typical usage (encoding usage patterns, inputs, output results), etc.

#### 4.2 Uncovering misuse

We now discuss detecting potential cases of misuse. The concept entails *misuse indicators*, which reflect certain criteria of tenant behaviour that may warrant some, or indeed, further attention. The implementation of an indicator entails an analysis of the relevant audit information sources to determine whether the particular criteria has been met. In practice, there are many possible types and forms of indicators, some generic and some context-specific. In this section we present the general concept to offer a way forward, using facial recognition as an illustrative example [10].

*Specific misuse indicators.* There will be situations in which providers will have some knowledge or foresight as to the particular risks of a particular service, or the forms that misuse might take. For instance, in facial recognition contexts, some concerns relate to population surveillance, personal privacy, and use by state services [3, 40].

Table 2 presents some specific examples of misuse indicators for vision and face-based AlaaS services and their category of risk. We now elaborate these (with reference to Fig. 2 and Table 1):

- *High request rate for face services*: If the tenant’s request rate for face services is high, it possibly indicates facial recognition deployment at-scale (e.g., population surveillance). Indeed, we see a rate limiting approach already employed by Microsoft [20]. Request rate patterns are easily derived from the transaction metadata already captured by platforms.
- *Large number of faces in an image/video*: Also concerned with surveillance, this indicator targets whether tenants appear to be analysing crowds by looking for many faces within the same input. We note Amazon AWS limits the number of faces detected in any image [2], though it is unclear if this is due to surveillance concerns.

**Table 2: Some examples of some AIaaS misuse indicators for object and facial recognition services.**

<i>Technical indicator for vision and face services</i>	<i>Potential implications</i>
High request rate for face detection	Population surveillance
Large number of faces in an image/video	Population surveillance
Large number of different faces are analysed	Population surveillance
Large number of identification attempts for particular individual(s)	Privacy threats to an individual
Detection of 'black-listed' objects	Controversial application

- *Large number of different faces are analysed*: A consistent use of a service to uncover different faces *over time* could also indicate a population surveillance scenario. This indicator could include considering aspects such as the frequency and volume of faces matched (i.e. numbers regarding input/output faces), to more sophisticated (and potentially more intrusive and legally-challenging) approaches, which might involve, for example, recording and analysing the raw input images or other details of the faces involved.
- *Large number of identification attempts for particular individual(s)*:<sup>2</sup> This concerns a tenant searching for a similar face across a large number of images coming from different contexts, suggesting the targeting and tracking of particular individual(s). Such an analysis might involve keeping records of the inputs/outputs to particular services, the results of processing, etc.
- *Detection of 'black-listed' objects*: This indicator looks for certain categories of items that might generally be of a type indicating a problematic application. For example, repeated detection of placards, or 'violence' could indicate that the technology is being used to screen protests. Moreover, this example illustrates the relevance of monitoring *across* different AIaaS services; e.g. detecting blacklisted objects and faces from the same inputs, such as placards (object detection) and multiple faces (facial recognition), may be revealing.

*Misuse discovery through patterns of behaviour.* There will also be situations where a specific risk might not have been foreseen, but where the usage pattern itself might indicate a need for further scrutiny.

This entails building profiles of AIaaS usage for and across particular services. Analysis might involve using, for example, anomaly detection methods [4] to indicate potentially problematic usage patterns. This is similar to how such methods are used in the security domain, e.g. for intrusion detection, where changes in network traffic can indicate that a system has been compromised [1].

This idea could be useful by indicating where the behaviour of a customer deviates from their normal usage pattern. For instance, if a customer transaction begins using a face recognition API far more extensively than previously, or integrates face detection with other services having not previously done so, it might warrant further investigation. Similarly, one can compare the usage profiles across customers. Assuming most customers (as organisations with responsibilities) are behaving appropriately, detecting a particular usage pattern that contrasts to that of others may warrant attention.

<sup>2</sup>Identification in this context is described as a one-to-many comparison [10], i.e. 'looking for' individual(s).

A further example concerns patterns indicating a model inversion attack [11], whereby tenants may be attempting to reverse-engineer the classification-criteria or training data from the model itself. Given the value of AIaaS is in the model, and the potential legal implications should models encode personal data (see [36]), this might be of particular concern for providers. Such attacks may require coordination from across a range of accounts, and there appears space for exploring the use of ML/anomaly detection towards detecting such situations.

## 5 PRACTICAL CHALLENGES & OPPORTUNITIES

So far we have introduced the concept of AIaaS monitoring as a means for indicating possible misuse. This section discusses some relevant legal and technical considerations, highlighting some interesting practical challenges and opportunities for future work.

### 5.1 Legal

AIaaS providers monitoring customers' use of their services and models may have legal consequences. We now explore these in relation to data protection law and intermediary liability, noting this is an area requiring further consideration. As such, rather than undertaking an in-depth analysis of the various legal issues and questions potentially arising from monitoring AIaaS, we highlight some considerations with a view to exploring these questions in greater detail in subsequent work. Note that, as well as the issues discussed below, further considerations are likely to arise in relation to the enforcement of contractual provisions and terms of service applying between AIaaS providers and their customers.

*Data protection.* Under the EU's General Data Protection Regulation (GDPR) [34] (and similar legislation in other jurisdictions), some of the audit information described above may be classed as *personal data* (i.e. any information relating to an identified or identifiable natural person [GDPR Art 4.1]). Depending on the specifics, this would likely include much information that we have categorised in Table 1 as 'transaction raw data', such as metadata (IP address), input parameters, and the information returned to clients, and some kinds of processing by-products. Note, however, that the operational metrics described earlier and some derived data are unlikely to be personal data, should they entail, for instance, aggregated or more general information about customers' use of services (e.g., regarding transactions, rates of service use, etc).

GDPR establishes a framework governing the *processing* of personal data (essentially any use or operation on personal data [GDPR Art 4.2]), with strict compliance requirements and potentially serious penalties for non-compliance. AIaaS providers who process

personal data in monitoring customers' use of their services, as in some of the monitoring methods this paper proposes, are likely to be data controllers [GDPR Art 4.7] for that monitoring – as they determine its means (how they are doing it) and its purposes (why they are doing it). This means they would then be subject to the GDPR's extensive compliance obligations [GDPR Art 5.2].

Providers may in practice be unable to determine which of the data that they are processing is personal data without first processing it. As a result, providers might best, in processing such kinds of audit information, take a precautionary approach that treats all of the data that has the potential to be personal as personal data. Moreover, GDPR recognises certain 'special categories' of personal data that are particularly sensitive [GDPR Art 9]. Given the inability to determine which data falls into one of these categories without processing it, cloud providers may need to treat large swathes of data for monitoring as if it was special category.

All personal data processing requires a basis in law [GDPR Arts 5, 6, 9]. The available bases for processing special category data are restricted [GDPR Art 9], and it's likely that only *explicit* consent of data subjects (individuals to whom personal data relates) would be available in these circumstances. It is not obvious how AlaaS providers could identify the data subjects to seek or obtain their explicit consent to the processing for monitoring without first processing the data in question. In circumstances where personal data is likely to be involved, cloud providers monitoring tenants' use of models would therefore face a dilemma. They will not in many cases be able to monitor without the explicit consent of any data subjects whose special category data will be processed in doing so. Without processing the data, they would be unlikely to be able to identify data subjects in order to obtain their explicit consent.

*Intermediary liability.* Many jurisdictions provide online platforms with protection from liability for any hosted content that might be illegal or unlawful in some way. In different jurisdictions, this protection may be given on a blanket basis, or may be qualified in some way or subject to certain conditions. In the EU, this is provided for in the E-Commerce Directive (ECD) [33]. Protection from liability is available to service providers who store information on behalf of users of their services by passive, technical means, so long as the provider does not have actual knowledge of any illegality [ECD Art 14, recital 42]. If the provider acquires such knowledge, they are obliged to remove the illegal information or activity expeditiously [ECD Art 14].

It is uncertain whether cloud providers offering AlaaS could avail of liability protections for hosting, due to the nature of the active role they play in providing that service. The current jurisprudence of the European Court of Justice indicates that, in some circumstances, service providers who take an active role may go beyond activities for which protection from liability is provided [24][25]. However, the Court's case law is currently unclear on precisely which kinds of activities would involve a provider taking such a role (although note that, as of December 2019, the Court is considering a case that may provide an opportunity to clarify the law in this area [23]).

Here, AlaaS providers are not simply offering a hosting facility by purely passive, technical means, but rather, offer access to sophisticated models that classify, predict, make decisions, etc. Providers

may therefore have some liability for their use. However, if they *can* ordinarily avail of that protection, actively monitoring tenants' activities could take them outside of that protection. In order to avoid legal repercussions, they would be obliged to expeditiously remove or prevent access to any content or activity that their monitoring determined to be illegal. Cloud providers might therefore prefer to 'play dumb' in order to avoid triggering any obligations and legal risks under the Directive. This is an interesting area requiring further investigation.

*Future.* Note, however, that in some jurisdictions, including the EU, there are proposals for reforming the law around the responsibilities of online platforms. Several of these propose to place greater responsibilities on platforms around the information they host or process (e.g. [9, 35]). This may put providers under obligations to monitor use of their models (which, given that they are providing access to powerful probabilistic models at scale and with a low barrier to entry, may not be an unreasonable development). If monitoring became explicitly required by law, that could potentially also allow them to circumvent the problems with obtaining explicit consent from data subjects. This is because GDPR provides a legal basis for processing special category data where doing so is necessary for reasons of substantial public interest and is undertaken on the basis of EU or Member State law that provides for sufficient safeguards for data subject rights [GDPR Art 9].

## 5.2 Technical

Monitoring for AlaaS misuse by platforms also raises a number of interesting technical challenges and opportunities for research.

A key area of research concerns the technical mechanisms to support oversight, which includes (i) technical monitoring architectures, and (ii) the analysis and development of misuse indicators. Some approaches may aim to operate generally, while others may be more context/application-specific. Work on intrusion detection and threat monitoring in networked contexts (§4.2), as well as work on provenance [30], appear relevant as possible starting points. Testing and ensuring that the indicators are meaningful and accurate is also an area for consideration.

Another aspect relates to performance. Service providers naturally seek to optimise the performance and reduce the resource requirements of the services that they offer. At the same time, monitoring entails overhead (e.g. processing, storage, etc.). As such, there are opportunities for research into quantifying the overheads and other impacts of various monitoring and detection processes. Note again that some aspects, such as request metadata, are likely already to be captured by providers, such as to assist with billing (§2.3), and so there may be ways to develop monitoring methods that leverage these to avoid adding substantial extra overheads.

Further, and in line with the above discussion, there may well be legal issues and liability implications from providers undertaking monitoring, and more generally, being more actively involved in policing the use of their services. The specifics of any potential increase in liability exposure will certainly depend on the circumstances. Such concerns also represent opportunities for technical research; examples include devising indicators and methods for monitoring that are more privacy and data-protection 'friendly'.

## 6 CONCLUSION

AIaaS makes accessible a wide range of sophisticated and scalable AI services for customers to integrate into their applications. These services act as application ‘building blocks’, which are potentially accessible by anyone. ‘Turnkey’ access to such services reduces the direct interactions with the provider, and so too the opportunities for oversight and intervention. It is therefore foreseeable that there will be situations in which AIaaS is used for controversial purposes, be it intentionally or otherwise.

While issues surrounding AI ethics, regulation and responsibility are currently the subject of much discussion, the potential for AIaaS misuse has had comparatively little attention. This is important, as given the complexity, data and skills required for ML undertakings, it is likely that AIaaS will become a dominant part of the technical infrastructure underpinning AI-driven applications. There appears a role for AIaaS providers to be more active in governing the use of their offerings, be it for reasons of reputation or responsibility – not least given the growing demands for greater tech-accountability.

In this paper, we used illustrative examples to highlight related issues, providing an initial conceptual overview of how providers might work towards ensuring more responsible use of their services. We also indicated the potential challenges and opportunities for moving forward, both from a legal and technical perspective.

In all, by presenting this concept, we seek to draw attention to AIaaS and its usage as an area warranting further consideration.

## ACKNOWLEDGMENTS

We acknowledge the financial support of the Engineering & Physical Sciences Research Council, University of Cambridge, Microsoft via the Microsoft Cloud Computing Research Centre, and Aviva.

## REFERENCES

- [1] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. 2016. A Survey of Network Anomaly Detection Techniques. *Journal of Network and Computer Applications* 60 (2016), 19–31.
- [2] Amazon. 2019. *Build an AI-driven Application*. <https://aws.amazon.com/machine-learning/ai-services/>
- [3] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitoff, Bobby Filar, et al. 2018. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228* (2018).
- [4] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.
- [5] J. Cobbe. 2019. Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making. *Legal Studies* (2019).
- [6] Kate Conger, Richard Fausset, and Serge F Kovaleski. 2019. *San Francisco Bans Facial Recognition Technology*. <https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html>
- [7] Sabbagh Dan. 2019. *Facial Recognition Technology Scrapped at King's Cross Site*. <https://www.theguardian.com/technology/2019/sep/02/facial-recognition-technology-scrapped-at-kings-cross-development>
- [8] Jared Dean. 2014. *Big data, data mining, and machine learning value creation for business leaders and practitioners* / Jared Dean. Wiley, Hoboken.
- [9] Department for Digital, Culture, Media & Sport. 2019. *Online Harms White Paper, CP 57*. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/793360/Online\\_Harms\\_White\\_Paper.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf)
- [10] European Union Agency for Fundamental Rights. 2019. *Facial Recognition Technology: Fundamental Rights Considerations in the Context of Law Enforcement*. <https://fra.europa.eu/en/publication/2019/facial-recognition>
- [11] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proc. ACM SIGSAC*. ACM, 1322–1333.
- [12] Google. 2019. *AI and machine learning products*. <https://cloud.google.com/products/ai/building-blocks/>
- [13] Google. 2019. *Our approach to facial recognition*. <https://ai.google/responsibilities/facial-recognition/>
- [14] Douglas Harris. 2018. Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke L. & Tech. Rev.* 17 (2018), 99.
- [15] Keiko Hashizume, Nobukazu Yoshioka, and Eduardo B Fernandez. 2013. Three Misuse Patterns for Cloud Computing. In *Security Engineering for Cloud Computing: Approaches and Tools*. IGI Global, 36–53.
- [16] Dirk Helbing. 2019. Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards Digital Enlightenment*. Springer, 47–72.
- [17] Vincent James. 2019. *Microsoft Denied Police Facial Recognition Tech Over Human Rights Concerns*. <https://www.theverge.com/2019/4/17/18411757/microsoft-facial-recognition-sales-refused-police-access>
- [18] Jens Lindemann. 2015. Towards Abuse Detection and Prevention in IaaS Cloud Computing. In *2015 10th International Conference on Availability, Reliability and Security*. IEEE, 211–217.
- [19] Bernard Marr. 2018. The AI Skills Crisis And How To Close The Gap. <https://www.forbes.com/sites/bernardmarr/2018/06/25/the-ai-skills-crisis-and-how-to-close-the-gap/>
- [20] Microsoft. 2019. *Cognitive Services*. <https://azure.microsoft.com/en-gb/services/cognitive-services>
- [21] Microsoft. 2019. *Six Principles For Developing and Deploying Facial Recognition Technology*. <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2018/12/MSFT-Principles-on-Facial-Recognition.pdf>
- [22] Arianne E Miller. 2018. Searching for Gaydar: Blind Spots in the Study of Sexual Orientation Perception. *Psychology & Sexuality* 9, 3 (2018), 188–203.
- [23] European Court of Justice. [n.d.]. LF v YouTube (C-682/18).
- [24] European Court of Justice. 2010. Google France SARL and Google Inc. v Louis Vuitton (C-236/08) ECLI:EU:C:2010:159.
- [25] European Court of Justice. 2011. L'Oréal SA and Others v eBay International AG and Others (C-324/09) ECLI:EU:C:2011:474.
- [26] Saeedeh Parsaefard, Iman Tabrizian, and Alberto Leon-Garcia. 2019. Artificial Intelligence as a Services (AI-aas) on Software-Defined Infrastructure. *arXiv:cs.LG/1907.05505*
- [27] Siani Pearson and Azzedine Benameur. 2010. Privacy, Security and Trust Issues Arising from Cloud Computing. In *Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CLOUDCOM '10)*. IEEE Computer Society, Washington, DC, USA, 693–702.
- [28] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. A Survey of Machine Learning for Big Data Processing. *Eurasip Journal on Advances in Signal Processing* 2016, 1 (2016).
- [29] Francesca Rossi. 2019. Building Trust in Artificial Intelligence. *Journal of international affairs* 72, 1 (2019), 127–134.
- [30] J. Singh, J. Cobbe, and C. Norval. 2019. Decision Provenance: Harnessing Data Flow for Accountable Systems. *IEEE Access* 7 (2019), 6562–6574. <https://doi.org/10.1109/ACCESS.2018.2887201>
- [31] Brad Smith. 2018. *Facial Recognition: It's Time for Action*. <https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/>
- [32] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. In *25th USENIX Security Symposium*. 601–618.
- [33] European Union. 2000. Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive) OJ L 178.
- [34] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119.
- [35] European Union. 2019. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC OJ L 130.
- [36] Michael Veale, Reuben Binns, and Lilian Edwards. 2018. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (2018), 20180083. <https://doi.org/10.1098/rsta.2018.0083>
- [37] Kent Walker. 2018. *AI for Social Good in Asia Pacific*. <https://www.blog.google/around-the-globe/google-asia/ai-social-good-asia-pacific/amp/>
- [38] Sarah Whittaker. 2018. What's The Big Deal? The Controversy on Google's AI and Pentagon Drones. <https://dronebelow.com/2018/03/08/whats-the-big-deal-the-controversy-on-googles-ai-and-pentagon-drones/>
- [39] Doffman Zak. 2019. *Hong Kong Exposes Both Sides Of China's Relentless Facial Recognition Machine*. <https://www.forbes.com/sites/zakdoffman/2019/08/26/hong-kong-exposes-both-sides-of-chinas-relentless-facial-recognition-machine/>
- [40] Yi Zeng, Enmeng Lu, Yinqian Sun, and Ruochen Tian. 2019. Responsible Facial Recognition and Beyond. *arXiv:1909.12935* (2019).
- [41] Lina Zhou, Shimei Pan, Jianwu Wang, and Athanasios V. Vasilakos. 2017. Machine Learning on Big Data: Opportunities and Challenges. *Neurocomputing* 237, December 2016 (2017), 350–361.