Towards Privacy Requirements for Collaborative Development of AI Applications

Vida Ahmadi Mehri*, Dragos Ilie† and Kurt Tutschku‡
Dept. of Computer Science and Engineering
Blekinge Institute of Technology, Karlskrona, Sweden
*vida.ahmadi.mehri@bth.se, †dragos.ilie@bth.se, ‡kurt.tutschku@bth.se

Abstract:
The use of data is essential for the capabilities of Data-driven Artificial intelligence (AI), Deep Learning and Big Data analysis techniques. The use of data, however, raises intrinsically the concern of the data privacy, in particular for the individuals that provide data. Hence, data privacy is considered as one of the main non-functional features of the Next Generation Internet. This paper describes the privacy challenges and requirements for collaborative AI application development. We investigate the constraints of using digital right management for supporting collaboration to address the privacy requirements in the regulation.

I. Introduction

Collaborative application development across organizations has become a major focus in Data-driven Artificial Intelligence (AI) system design when aiming at sophisticated AI applications[1], [2]. This collaboration process builds on specialisation in AI engineering and on re-useable AI objects, e.g. data set or Deep Learning models. These objects have been gathered or developed by third-parties not designing the final application. The advantages of the process are a possible significant reductions of development cost and time and an option for engineering for higher AI performance. The appealing features are evidenced by the development of AI pipelines [3], open source machine learning and data visualization tools such as Orange [4] and the emerge of data marketplaces [5], [6].

This collaborative approach, however, comes at a cost. It imposes at least three fundamental challenges on the design process. First, the use of data intrinsically raises data privacy concerns. These doubts become even deeper regarding the feature of data set being shared. Second, Data-driven AI aims at identifying unknown relationships within the information. However, when using typical privacy enforcing mechanism such anonymization technique or restriction in data collection, then it cannot be excluded that inherent relationships within the data sets are not captured or deleted. As a result, such data sets are becoming useless. While privacy concepts are of high value for specific applications, they might impact the usability of AI objects in general and a dilemma for the general concept of collaboration based on re-usability arises. Third, the re-use of AI objects requires trust among the developers and users if new forms of collaboration are applied. This trust ranges from obeying licences between developers to permitting governance as required by societies and individuals on the use of the AI objects, e.g. enabling GDPR or GDPR-like concepts on the use of data and AI objects in Europe. This paper aims to address these fundamental challenges by giving the insight to the privacy requirements in collaborative AI development. It will provide an initial taxonomy of privacy and Digital Rights Management (DRM) and the threats against objects in the AI pipeline. The paper summarise the GDPR act and its potential implications for Bonseyes-like AI marketplaces and describe potential ways of violating the DRM associated with the artefacts.

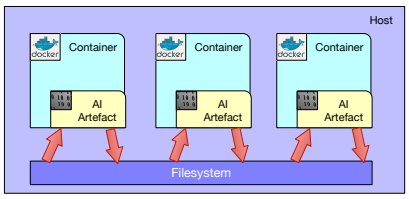

Fig. 1. Original Bonseyes pipeline

II. Collaborative AI Application Development

The purpose of data-driven AI is to analyse collected data in a smart way and come up with useful predictions about future data or provide new insights about existing data. However, in order to achieve good results it is necessary to carefully prepare the data (*e. g.*, remove noise) and trim the algorithm parameters. The output of the algorithm, analytical processing of data, is a model which is used for predictions.

A model needs to be *deployed* at a location where it can process input data. The location can vary between a powerful cloud computing environment to resource-constrained IoT devices, depending on the intended application.

From a business perspective, we are witnessing the emergence of stakeholders that can provide access to high-quality data or algorithms. The co-dependency between data and algorithms in the AI workflow model suggests that collaboration between various stakeholders is required for developing complex, high-performant AI applications. To this end, the Bonseyes project is designing a AI marketplace that will enable such collaboration while maintaining privacy and enforcing DRM.

Bonseyes uses an Agile methodology for developing the marketplace. The current implementation of the AI workflow model is shown in Fig. 1 and is referred to as an *AI pipeline*. The light-blue rectangles in the figure are Docker containers. Inside the Docker container resides an *AI artefact*. The artefact can be pure data, an interface to a data source or an algorithm used in the AI workflow.

The developers retrieve the required AI artefacts from a Docker repository and assemble the pipeline on their local system. The red arrows represent data transfers to and from

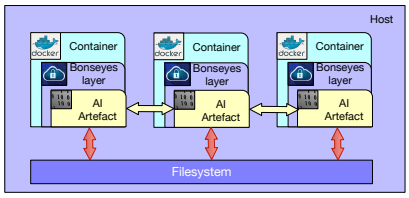

Fig. 2. Improved Bonseyes pipeline

storage (currently a regular filesystem). The transfers are needed to bring data in the pipeline and to save output from intermediate stages and the result. The data may include confidential information, hence the red color.

Fig 2 shows the next implementation of the AI pipeline, where the AI artefact is wrapped inside a *Bonseyes Layer(BL)*. The purpose of the *BL* is to enable secure direct artefact communication without needing to store intermediate data. It also provides DRM mechanisms to control access to artefacts and enforce license management policies. The need for DRM mechanisms is driven both by business motives as well as legal requirements.

The new implementation of the AI pipeline supports also a distributed model, as shown in Fig 3, which allows various components of the pipeline to execute on different hosts. This model requires that the *BL* is extended to the hosts in order to facilitate workflow distribution and is the model considered for the reminder of the paper.

III. GDPR PRIVACY REQUIREMENTS

General Data Protection Regulation(GDPR) is the uniform approach towards all EU countries to provide the protection of a natural person while processing the individual's personal data. It will come in force in May 2018 as a replacement of European data protection Directive (EU Directive 95/46/EC) to give the EU citizens better control over their personal information. GDPR expands the notion of personal data to photos and audios, financial transactions, posts on social medias, device identifiers, location data, users login credential, and browsing history, as well as genetic information. The new regulation applies in a wide territorial scope and it includes all countries (EU or non-EU) that process the personal data of EU residents.

GDPR relies on six main principles for processing personal data namely a) lawfulness, fairness, and transparency; b) purpose limitation; c) data minimization; d) accuracy; e) storage limitation; f) confidentiality and integrity [7].

GDPR limits the collection and storing of identification information of data subject up to the minimum necessary required in order to safeguard data subjects' rights and freedom.

The new regulation expand the rights of data subjects and lists the detail obligations and responsibilities for two key entities: controllers and processors. The *controller*, who "determines the purposes and means of the processing of personal data", is responsible for implementing technical and organisational measurement to ensure the processing of data is performed as described in the regulation. The organisational measurement might be to assign a data protection officer, to do data protection impact assessment and risk mitigation plan, as well as technical measurement such as implementing pseudonymization and data minimization. A controller can delegate the processing of personal data to a *processor*, who "processes personal data on behalf of controller" therefore it is the responsibility of the controller to select processors who guarantee the implementation of appropriate technical and organisational measures that meet the regulation requirements and ensure the protection of data subjects [7].

A. GDPR impact on AI pipelines

From a business point of view, the AI marketplace enables *artefact consumers* to obtain *licensed* access to AI artefacts owned by *artefact providers*. An entity can play both roles: it can be a provider for artefacts it owns, and it can consume artefacts from others. An artefact is coupled together with a license that specifies the terms of use for the artefact. Since the license is viewed as a legal document it must be anchored into existing laws and regulations (*e. g.*, GDPR) in order to be effective. A license can be encoded into a digital license to enable computer-based license management systems to monitor, detect and prevent license violations.

As described in Section II, AI pipelines are setup in order to achieve specific goals such as to make predictions or obtain new insights from existing data. Since the purpose of the pipeline is determined by the entity setting up, it is reasonable to conclude that the entity is acting as controller in the GDPR sense and thus is responsible for complying with the GDPR. When viewed in the context of an AI marketplace, individual AI artefacts in the pipeline act as generic building blocks and should not be aware of the pipeline's higher purpose. It is thus reasonable to assume that artefact providers act as GDPR processors. However, some providers may implement the functionality of an artefact by chaining together several other artefacts, where each element of the chain consume the output of the previous element and provides input to the next. The determination of whether the provider of a chained artefact acts as controller or processor does not seem so clear cut in these cases.

The distinction between controllers and processors becomes even blurrier when the AI artefact contains a data source or pure data. The artefact provider could have collected and processed private data from its data subjects for specific purposes, thus being itself a controller according to GDPR. The provider may view the pipeline owner as a processor or they may establish a joint controller relationship [7].

This discussion highlights some of the immediate difficulties in attributing legal responsibilities to providers and consumers. It is not quite clear at the moment how an AI marketplace could *automatically* assign correctly (in a legal sense) the controller/processor roles to various entities participating in a pipeline. We believe that a tractable initial approach is to determine a baseline of requirements (legal and technical) for processors. All consumers and providers in the marketplace would be required to fulfill these constrains. This would enable controllers to delegate data processing to any processor.

IV. DRM AND PRIVACY REQUIREMENTS FOR AI DEVELOPMENT

In general, we identify the following general types of usage constraints for an artefact:

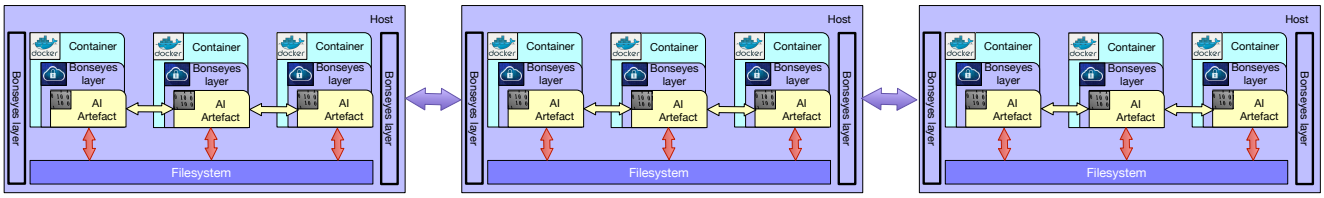

Fig. 3. Distributed Bonseyes pipeline

i) **Validity** is the duration of a license that called the license *validity period*. Additionally, the validity may be also constrained by number of times the artefact can be executed, a so-called n-times use price model. Also, a license should be revokable if users breach the license agreement or to allow users to interrupt an automatic subscription renewal.

ii) **Beneficiary** constraint specifies the identity of the *license user*. This can be an end-user license or a group license that allows an organisation to allocate and revoke licenses to its members according to own policies.

iii) **Purpose** constraint defines the *license scope* (*e. g.*, commercial, educational, or personal). This can even be tied to specific types of licenses (*e. g.*, GPL) that govern how derivative work (*e. g.*, AI applications) may itself be licensed. In addition, legislation, such as the GDPR, may stipulate how data must be collected, stored, shared and processed. The usage purpose constraint must be able to incorporate this type of law-derived policies.

iv) **Location** constraints define where the artefact can be utilised. This constraint can be either topological or geographic. The topological location defines the networked hosts that are allowed to use the artefact. The simplest topological constraint restrict artefact usage to the set of Bonseyes authorised hosts. However, artefacts can also be constrained to *virtual premises*, which are smaller subsets[8]. The geographical location defines where in the world the artefact can be used or is prohibited from being used. This allows hosts and artefact to comply with local laws and regulations, such as EU Data Protection Directive and its successor, GDPR.

v) **Peering** constraint regulates which entities the artefact is allowed to interact with. The peers can be the successor or predecessor artefact in a pipeline or entities that monitor and control the operation of the artefact.

A license management system must protect the interests of both artefact consumers and providers, according to license constraints. This means that consumers are not prevented from using the artefact while the license is valid, but also that they cannot continue to use the license if it expires, is revoked for a legitimate reason or if the artefact is used for a purpose or at a location prohibited by the license. Similarly, the providers must be prevented from blocking access to an artefact for consumers with a valid license, while at the same time retain the possibility to revoke a license when its terms are breached. The collaborative nature of the AI marketplace raises some interesting challenges in satisfying these requirements.

To ease the threat analysis, we will consider two classes of misbehaving users: regular users and malicious users. Regular users can be consumers that try using the artefact in conflict to the license constraints, or providers that attempt to prevent consumers from using an artefact although a valid license exists. Common for this case is that the users are either unaware that they are breaching the license agreement, or they are aware but donot employ any advanced means (*e. g.*, reverse engineering or code injection) to achieve their goals. Therefore, we consider these potential attacks as *simple threats*. On the other hand, the malicious users are assumed to be skilled attackers that may insert exploits into the AI pipeline or instrument the artefact's hosts in order to bypass the license and obtain unfettered access to the artefact of interest. We consider these to be *advanced threats*. Provider attempts to fraudulently prevent consumer with valid licenses from using artefact belong also to the advanced threats category.

A. Simple threats

We begin by considering two obvious threats:

T-1: User can modify license contents to bypass usage constraints

T-2: AI marketplace repudiates the issuance of a license

It is assumed that the license contents are stored inside the BL that encapsulates the AI artefact. An additional assumption is that the license contents as well as the AI artefact are digitally signed by the marketplace. The signature enables the BL to detected fraudulent changes to the license contents and asserts the origin of the data thus preventing the marketplace from repudiating an issued license. We assume the signature algorithm itself (*e. g.*, RCA-PSS, DSA or ECDSA) when used with a reasonable key length is infeasible to break. This counteracts threat T-1 and T-2.

A *license validity period* begins on a specific start date and stops at an expiration date. The expiration date can be left undefined by the licensor if perpetual validity is desired. Additionally, if an n-times use pricing model is used, the validity can be further restricted by the number of times the artefact is executed. A license can be revoked before the expiration date for reasons mentioned previously in this paper. To determine if a license is valid, the Bonseyes layer is dependent on having access to a time source, such as a Network Time Protocol (NTP) server, and to a license revocation database. We consider threats against the integrity or availability of the time source and the license revocation database. More specifically, we require security mechanisms are put in place to protect against:

T-3: Blocking communication with the time source

T-4: Blocking communication the license revocation server

T-5: Spoofing a time source

T-6: Rolling back the time on the artefact host to use artefact past expiration date

The *license beneficiary* uniquely identifies the licensed consumer. This information is determined at the time the license is acquired. Its presence in the license enables accounting and usage tracking. Since the license and AI artefact are protected by a digital signature, it is assumed that no simple threats exists against the integrity of this information. However, it is important to protect against the execution of an artefact by an unlicensed user (*e. g.*, in the case a pirate copy is made).

T-7: Artefact execution by unlicensed user

License purpose is metadata encoding the permitted use scope for the artefact. Hosts in a virtual premise have similar metadata configured for their BL. The artefact BL and the host BL each expose their use scope. Thus, the artefact can decide if a specific host meets its purpose constraint and the host can decide if it can allow artefact execution. For example, an educational virtual premise may refuse to execute commercial artefacts, or an artefact can avoid running on hosts that does provide adequate privacy protection (*e. g.*, according to GDPR). This constraint faces the following threats:

T-8: Host breach against license purpose

T-9: Artefact breach against license purpose

We foresee the following topological location threats:

T-10: Artefact execution outside a virtual premise (*i. e.*, host lacking a valid BL)

T-11: Artefact execution on wrong virtual premise

These threats are similar to the T-8, but they are more generic in scope. Virtual premises are host pools for AI pipelines with different constraints. The license purposes and performance are such a constraints.

T-12: Artefact execution outside allowable geographic region.

Threat T-12 is a special version of T-11. In this case, we consider mobile virtual premises, a host or a set of hosts that can move or be transported geographically. The issue we try to capture is that when geographic boundaries are being crossed (like country or state borders) different jurisdiction may apply to computation and data.

B. Advanced threats

In this section we consider threats from skilled malicious users. The assumption is that this type of users are able to modify both the artefact BL code as well as the host BL. Also, they are legitimate AI marketplace users and can, for example, obtain valid host licenses. Thus, these attackers have the capability to circumvent any authorization and authentication mechanisms running on systems that are under their control. Thus, the defence mechanism for T-1 to T-12 are defeated on all virtual premises under the control of the attacker. In particular, it means that license management mechanisms can be bypassed and AI artefacts obtained from the marketplace cannot impose any limitation on how they are used. The attacker can also extract the actual algorithms or data encapsulated by the artefact and make it available outside the marketplace. Even if artefacts employ encryption as a copy-protection scheme, they cannot be protected against this type of attacks. The crux of the problem is that for an artefact to be useful, it needs to be decrypted at some point. The attackers can instrument the hosts under their control using a tool such as Intel PIN[1] and locate the decrypted artefact payload. This type of approach was used to successfully defeat movie DRM [9].

Although the outlook for defending against advanced threats looks quite gloom, there are two complementary approaches that can raise the difficulty of mounting a successful attack. The first approach aims at providing a trusted computing environment for license management mechanisms to execute, in particular one in which the integrity of host and artefact BL are protected.

Our second proposed approach for providing a trusted environment is to remove the hosts from the control of a potential attacker. In particular, we envision a scheme, similar to [10], where the actual implementation of the AI pipeline is done by a third party, such as a cloud provider. The assumption is that the cloud provider's main business interest is in selling computation and storage and thus the provider has no incentive to engage in malicious activities agains the AI marketplace and its users.

V. Conclusion

The paper describes the privacy requirements for enabling collaborative concept in AI application development. It investigates the DRM constraints to address the privacy requirements by describing the possible threats. The proposed solution for these threats will be implemented in future work.

Acknowledgment: This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 732204 (Bonseyes). This work is supported by the Swiss State Secretariat for Education Research and Innovation (SERI) under contract number 16.0159. The opinions expressed and arguments employed herein do not necessarily reflect the official views of these funding bodies.

References

[1] I. Stoica, D. Song, R. Ada Popa, D. Patterson, M. W Mahoney, R. Katz, A. D Joseph, M. Jordan, J. M Hellerstein, J. E Gonzalez, et al. A berkeley view of systems challenges for AI. *arXiv preprint arXiv:1712.05855*, 2017.

[2] T. Llewellynn, M Fernández-Carrobles, O. Deniz, S. Fricker, A. Storkey, N. Pazos, G. Velikic, K. Leufgen, R. Dahyot, S. Koller, et al. BONSEYES: platform for open development of systems of artificial intelligence. In *Proceedings of the Computing Frontiers Conference*, pages 299–304. ACM, 2017.

[3] PipelineAI - Home. available at https://pipeline.ai/.

[4] Orange Data Mining Fruitful & Fun. available at https://orange.biolab.si/.

[5] Data Market Austria. available at https://datamarket.at/.

[6] Industrial Data Space e.V. available at http://www.industrialdataspace.org/.

[7] European Parliament and Council of the European Union. 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). In *Official Journal of the European Union*, volume 59. May 2016. ISSN: 1977-0677.

[8] V. Mehri and K. Tutschku. Flexible privacy and high trust in the next generation internet: The use case of a cloud-based marketplace for AI. In *Proceedings of SNCNW*, Halmstad University, Sweden, 2017.

[9] R. Wang, Y. Shoshitaishvili, C. Kruegel, and G. Vigna. Steal This Movie: Automatically Bypassing DRM Protection in Streaming Media Services. In *Proceedings of 22nd USENIX Security Symposium*, pages 687–702, Washington DC, USA, 2013.

[10] Ronald Petrlic. Privacy-preserving digital rights management in a trusted cloud environment. In *Proceedings of TrustCom*, Liverpool, England, UK, 2012.

[1] https://software.intel.com/en-us/articles/pin-a-dynamic-binary-instrumentation-too