PERSONAL DATA PROTECTION COMMISSION SINGAPORE

PRACTICAL GUIDANCE TO GRAB

Background

1. Grab is implementing a Proof of Concept (POC) to automate the classification of Grab customers' personal data as follows:

a) **Tagging.** Large Language Models (LLM) will be used to automate the tagging of data fields containing personally identifiable information (PII), such as the following:
   i. [Personal.ID] – refers to external identification numbers that can be used to uniquely identify a person (e.g., "NRIC", "FIN", "License Plate")
   ii. [Personal.Name] – refers to name or user name of a person
   iii. [Personal.contact_info] – refers to contact information of a person (e.g., "email", "phone", "address", "social media")
   iv. [Geo.Geohash] – refers to a public domain geocode system that encodes a geographical location (latitude and longitude) into a short string of letters and digits
   v. [Personal.Traits] – e.g., "Gender", "Age", "Nationality"

b) **Classification.** Rules will be used to recommend the classification of the data records, based on the tags provided by the LLM. Specifically, data will be classified "Restricted" if it contains any PII (direct identifier)[1], or at least three (3) PII (indirect identifiers)[2].

c) **Verification.** Data owners will then be notified to verify if the data is classified correctly.

2. Upon data classification, data protection measures are also put in place to ensure compliance with Grab's data policies for specific data records, including the following:

a) **Nullification of data fields[3].** For data records that have reached the respective retention periods specified in Grab's data retention policies, data owners will nullify the data fields containing any PII (direct identifiers) and/or PII (indirect identifiers).

b) **Data access controls.** To ensure that Grab employees do not have access to data fields beyond what their roles require, there will be data access controls to ensure that Restricted data can only be accessed by data owners and other employees ("users") on a need-to-know basis, with approval from the user's manager and data owner. Users who have default access to hashed Restricted data are from a controlled group of Grab's Technology Full Time Employees (FTE), who belong to a central technology team that supports various business units/entities within Grab to generate data/business insights.

[1] E.g., Data fields which have been tagged [Personal.ID], [Personal.Name], [Personal.contact_info]
[2] E.g., Data fields which have been tagged [Geo.Geohash], [Personal.Traits].
[3] "Nullification" refers to the removal of data values in a column of data.

c) **Hashing[4] of PII data fields in Restricted data.** Users without permission to access Restricted data will only be able to access the data records in a hashed format (i.e., data columns with PII (direct and indirect identifiers) will be hashed). For data that is Unrestricted (i.e., data records with fewer than 3 PII (indirect identifiers) and without any PII (direct identifiers)), users will be able to access all data fields within each record.

3. Grab sought Practical Guidance (Guidance) from the Personal Data Protection Commission (PDPC) on the following:

a) Whether data records that contain hashed PII fields **(see para 2(c))** constitute personal data under the PDPA.

b) Whether LLM can be relied upon to do data classification.

PDPC's assessment

4. PDPC's guidance set out below focuses on the following:

a) Whether the hashed data and nullified data constitute anonymised data[5], for which the Data Protection Provisions under the PDPA do not apply;

b) Potential data protection risks arising from the use of LLM to classify data; and

c) Whether there are applicable exceptions to consent that Grab may rely on for the use and/or disclosure of customer personal data for data analytics and generation of business insights.

Whether the hashed data and nullified data constitute anonymised data

5. PDPC notes that Grab has implemented good data protection practices that combined technical and process controls. In particular, PDPC recognises that anonymisation techniques[6] had been applied to enhance protection of personal data, while enabling the use of data for insights and data innovation (e.g., data analytics, data modelling). PDPC treats anonymisation as a risk-based process which includes applying both anonymisation techniques and safeguards (i.e., technical, process, administrative) to prevent re-identification. In determining whether personal data is anonymised, PDPC will take into account the following[7]:

a) Whether all direct identifiers have been removed;

b) Whether indirect identifiers that can be used to re-identify individuals when matched with publicly available or proprietary information that the data recipient has access to have been altered or removed;

---

[4] "Hashing" refers to the conversion of data into another value (e.g., fixed length string of letters and numbers) using a hash function/algorithm.

[5] Anonymisation refers to the process of converting personal data into data that cannot be used to identify any particular individual.

[6] Refer to PDPC's Guide to Basic Anonymisation

[7] Refer Chapter 3 on Anonymisation in PDPC's Advisory Guidelines on the PDPA for Selected Topics

c) Whether there are additional safeguards implemented to restrict access and use of anonymised data to reduce the risks of re-identification (e.g., organisational structures, policies, processes); and

d) Whether there are periodic reviews conducted to assess adequacy of anonymisation techniques and risk management controls in relation to current state of technology, robustness of organisational, legal, processes and other non-technical measures to manage the risks of re-identification.

6. Based on the above, where both direct and indirect identifiers in a data record is nullified (i.e. removed), re-identification risk will be low. As such, PDPC would consider the data record to be anonymised.

7. For Restricted data, PDPC notes that hashing will be performed on both direct and indirect identifiers in each data record. While hashes are cryptographically generated strings that serve as irreversible one-to-one representations of the data that was hashed, proper safeguards should be implemented to prevent attackers from identifying individuals through inferences from pre-computed tables. Grab should ensure that the hashes generated are reasonably strong (e.g., by using industry-standard algorithms and incorporating a salt) to protect the data, particularly in the case of direct identifiers that follow pre-determined formats such as National IDs.

8. For Unrestricted data, PDPC notes that data records with fewer than 3 PII (indirect Identifiers) will not be hashed and can be accessed in the clear. While no direct identifiers are included in such data records, attackers/unauthorised parties may still re-identify individuals by querying and merging multiple records belonging to an individual and gaining access to data records with indirect identifiers in the clear. Such Unrestricted data would not be considered as anonymised. In particular, Grab will need to comply with the Protection Obligation by putting in place proper access controls and safeguards to protect such data, such as:

a) Monitoring of queries made, and/or random sampling/audit on persistent querying of data records with fewer than 3 PII (indirect identifiers);

b) Review of access policies (e.g., criteria for granting Restricted / Unrestricted user access rights, duration of access); and

c) Periodic review of user accounts to ensure that access policies are implemented (e.g., all the accounts are active and the rights assigned are in compliance with access policies, timely removal of user accounts when a user has left the organisation or update the user's rights when he/she has changed his/her role within the organisation).

Potential data protection risks arising from the use of LLM to classify data

9. PDPC recognises that data classification can be an effective tool to aid organisations in managing their data protection risks (e.g., by tailoring different sets of data protection measures/governance controls based on the data categories as defined by the organisation's internal classification policies).

10. In Grab's case, a combination of LLM tagging of data fields and rules-based classification is deployed to determine whether a data record qualifies as Restricted data. PDPC notes that there is a possibility that the LLM may not perform the tagging as intended, resulting in a downgrade in classification from Restricted to Unrestricted. This may increase

the risk of "unauthorised user access" where a user gains access to supposedly Restricted data in the clear (when the PII within the data record should have been hashed). To address and to reduce the likelihood of inaccurate data tagging and classification, Grab has put in place safeguards to monitor the accuracy of LLM (e.g., periodically using hard coded business rules to counter check the tagging and classification of randomly selected data records), and to ensure that there are additional checks (e.g., manual verification) on the classification output.

Applicable exceptions to consent under the PDPA

11. Where relevant, Grab may consider relying on the following PDPA's exceptions[8] to the Consent Obligation when using personal data:

a) **Business improvement exception** is likely to apply where Grab's use of personal data is to generate insights to improve or develop new goods or services, or to better understand customer preferences and behaviour etc. To rely on the exception, Grab will need to ensure that the purpose cannot be reasonably achieved without using the personal data in an individually identifiable form, and that a reasonable person would consider the use of personal data for such purpose appropriate in the circumstances. Grab may also rely on the business improvement exception to share personal data, without consent, between entities belonging to a group of companies[9] (e.g., Grab group).

b) **Legitimate interests exception** is likely to apply where Grab's use and/or disclosure of customers' personal data is for purposes such as fraud detection and preventing misuse of Grab's services. To rely on this exception, Grab will need to assess the adverse effect of the use and/or disclosure of personal data and ensure that the legitimate interests (i.e., benefits to Grab, other organisations, or wider segment of the public) in doing so outweigh any adverse effect on the individual.

END OF DOCUMENT

[8] Refer to Chapter 12 of PDPC's Key Concepts Advisory Guidelines for more details on business improvement and legitimate interests exceptions.

[9] "Group of companies" refers to related corporations within the meaning of the Companies Act (Cap. 50).