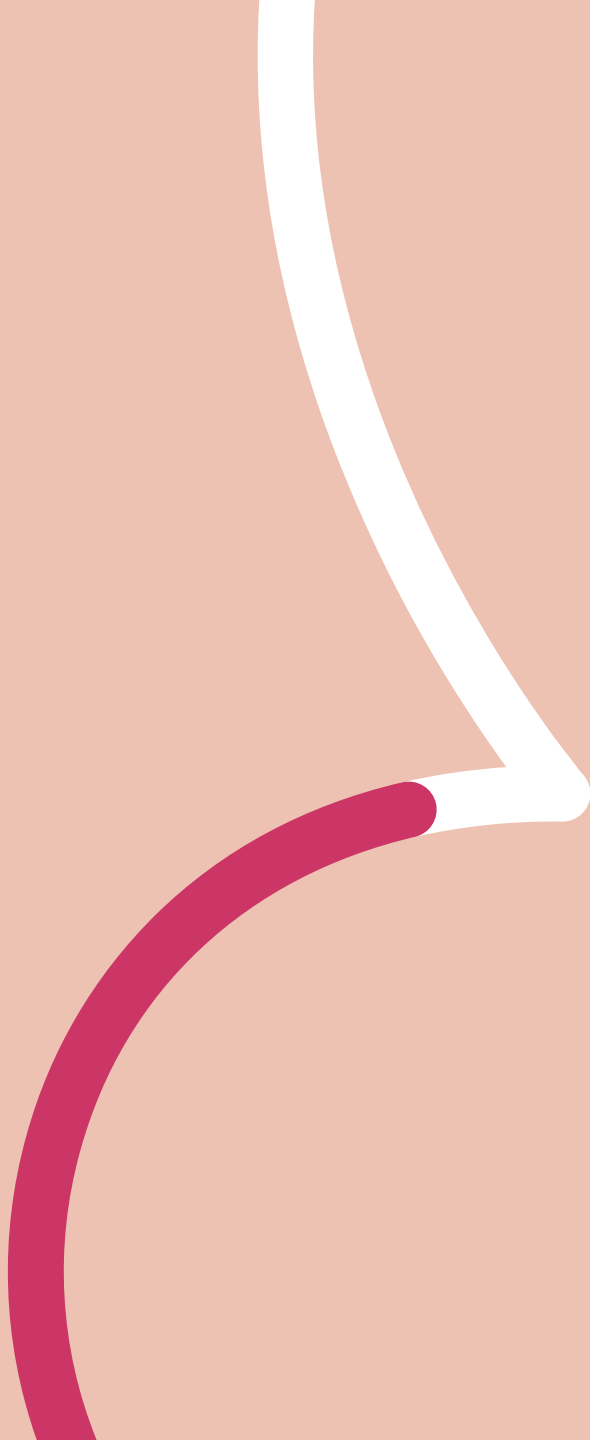


Bird & Bird

Gen AI: use of personal data for training AI systems

Gabriel Voisin (Privacy & Data Protection Partner)

4 March 2025



1

Can you rely on
legitimate interests to
use personal data for
training AI systems?

Data protection authorities' positions

No, according to the Dutch DPA

The strongest view **against** LI comes from the Dutch DPA: "*scraping will almost always be a violation of the General Data Protection Regulation*"

However, the AP's view is based on their longstanding highly restrictive approach to LI (purely commercial interests cannot be legitimate interests), which was recently proven wrong in the KNLTB case (see more on the next slide).



Update:

On 4 October 2024, the Court of Justice of the European Union [pronunciation](#)  done on the [Basis of legitimate interest](#). As a result, the AP's earlier explanation of this basis is no longer relevant. This may also have consequences for this guide on scraping. The DPA is currently considering whether the guideline needs to be revised.

Yes, according to the English, French and Belgian DPAs

Both ICO and CNIL have issued guidance that recognise it is possible to rely on LI for AI training, subject to the completion of a successful Legitimate Interest Assessment (LIA).

There has also been a Belgian decision where the APD concluded that a bank could rely on this basis for using payment transactions to train an AI model for personalised discounts service. The bank distinguished between personalised information based on the data subject's consent, that they could withdraw, and the model building, which was based on legitimate interest with the data subject's right to object. The Belgian DPA found that the processing to train the model did meet the three-part legitimate interest test.

CJEU's insights

Google v AEPD and Costeja case (C-131/12)

As provided by the AG in its opinion supported by the Court, *"it seems obvious that provision of internet search engine services pursues as such legitimate interests (...) namely (i) making information more easily accessible for internet users; (ii) rendering dissemination of the information uploaded on the internet more effective; and (iii) enabling various information society services supplied by the internet search engine service provider that are ancillary to the search engine, such as the provision of keyword advertising."*

KNLTB case (C-621/22)

In this case, the CJEU emphasised Recital 47 GDPR which confirms a legitimate interest does not need to be established in law and a wide range of interests - **including a purely commercial interest** - is, in principle, capable of being regarded as legitimate.

The EDPB is also hinting that it is possible, e.g.

EDPB Opinion 28/2024 on AI models

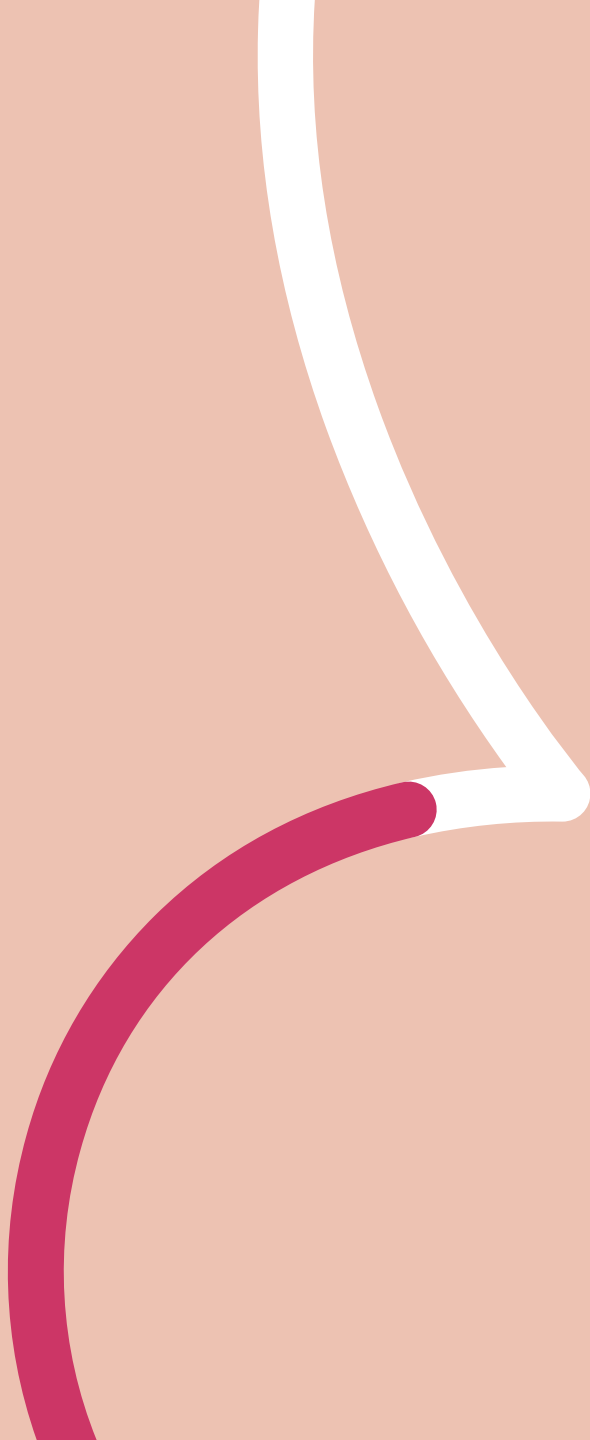
In this opinion, the EDPB confirms that controllers may rely on legitimate interests as a legal basis for processing personal data during the development and deployment of AI models if a thorough LIA is conducted and appropriate safeguards are in place.

It also points out that the following examples may constitute a legitimate interest in the context of AI models: (i) developing the service of a conversational agent to assist users; (ii) developing an AI system to detect fraudulent content or behaviour or (iii) improving threat detection in an information system.

EDPB draft Guidelines 01/2025 on Pseudonymisation

The recent draft opinion on pseudonymisation argues that the *"risk reduction resulting from pseudonymisation may enable controllers to rely on legitimate interests under Art. 6(1)(f) GDPR as the legal basis for their processing provided they meet the other requirements"*.

Next rendez-vous? The upcoming publication of the EDPB guidelines on web scrapping



2

What are the main challenges? Few examples...

Challenge 1: SCD

The background is a solid light purple color. On the right side, there are two thick, wavy lines. The lower line is a light pink color, starting from the bottom left and curving upwards and to the right. The upper line is a darker red color, starting from the top right and curving downwards and to the left, meeting the pink line.

How to deal with special category data (SCD)?

- The current approach of regulators advocates filtering and - in some cases - suggest it might reduce the need to comply with Art. 9:
 - CNIL: *“If, despite the measures taken, the organisation processes incidentally and residually sensitive data that it had not sought to collect, it is not considered illegal (...). [If] the organisation comes to know that it is processing sensitive data, it has to proceed, as far as possible, to its immediate and automated deletion.”*
 - The Dutch DPA makes a similar point but takes the view that this is a matter for the court.
 - Both are predicated on the **Google v. CNIL (C-507/17)**, which determined that search engine operators could comply reactively with de-referencing requests on the basis that their compliance requirements, inc. under Articles GDPR 9 & 10, applied 'in the context of their responsibilities, powers and possibilities'.
- This might help with 'incidental' SCD, but deliberate training on SCD – including to train on representative datasets – is likely to be more challenging and require an Article 9 GDPR condition

Possible options to deal with SCD

"Manifestly made public by the data subject" (Art. 9(2)(e) GDPR)

This is a high bar as illustrated by the CJEU in the *Meta v Bundeskartellamt* (C-252/21) where it was held that: *"where he or she enters information into such websites or apps or where he or she clicks or taps on buttons integrated into those sites and apps, such as the 'Like' or 'Share' [...], that user manifestly makes public, [...] the data thus entered [...] only in the circumstance where he or she has **explicitly made the choice** beforehand, as the case may be on the basis of **individual settings** selected with **full knowledge of the facts**, to make the data relating to him or her publicly accessible to an unlimited number of persons"*

"Substantial public interest" (Art. 9(2)(e) GDPR) combined for e.g. with Art. 10(5) EU AI Act

This article allows providers of high-risk AI systems to process special category data *"to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction in relation to the high-risk AI systems"*

BUT:

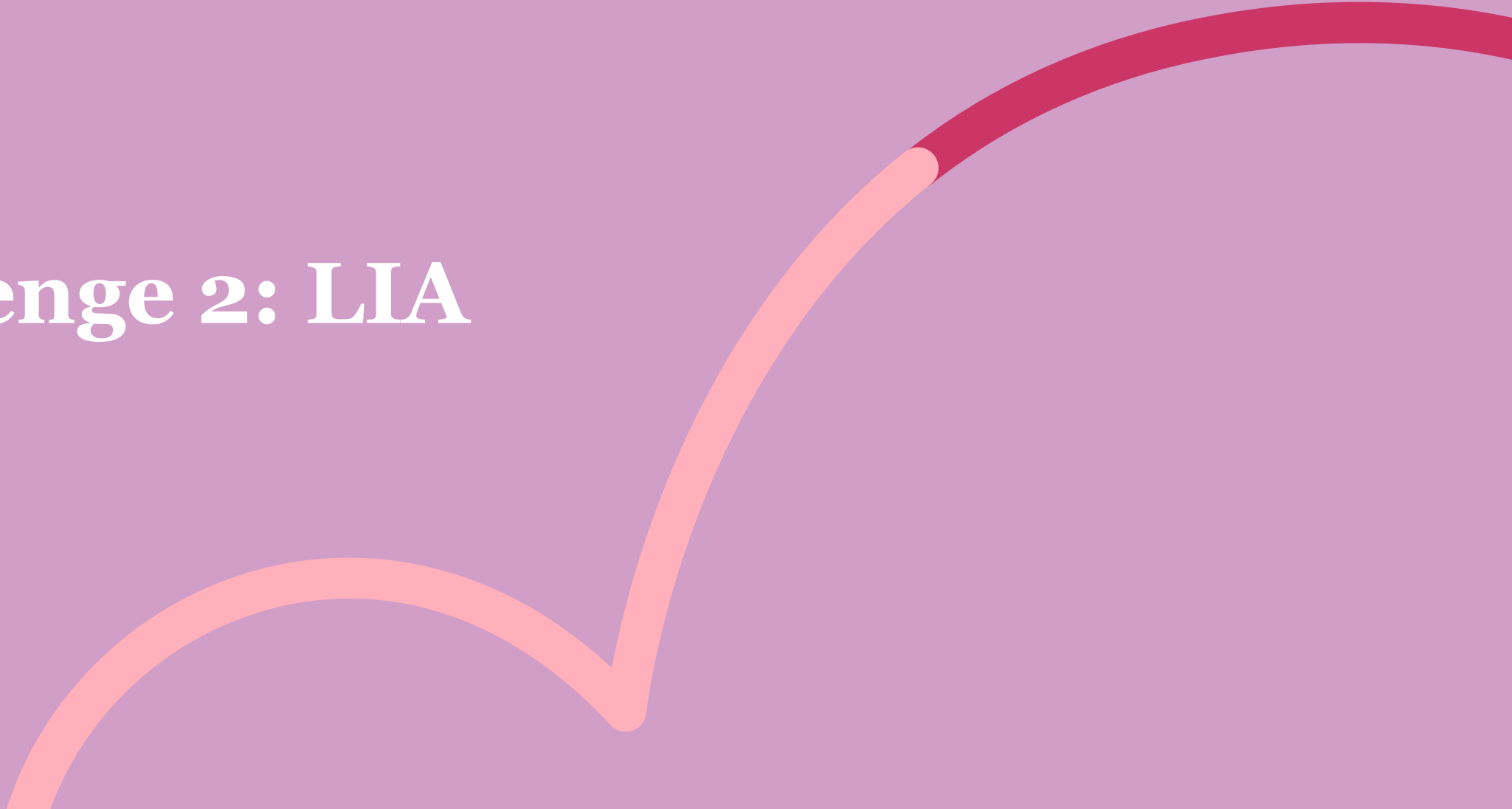
- this only applies to providers of high-risk AI systems.
- it does not apply to deployers or providers of general-purpose AI models with systemic risk

Other options to deal with SCD?



**call for
contribution**

Challenge 2: LIA

The background is a solid light purple. On the right side, there are two thick, wavy lines. The lower line is a light pink color, starting from the bottom left and curving upwards and to the right. The upper line is a darker red color, starting from the top right and curving downwards and to the left, meeting the pink line.

First question first...



Have you done one
(or more)?

How do you do a robust LIA?

1 The interest must be "legitimate"

It is **presumed legitimate** when:

- manifestly lawful under the law;
- determined in a sufficiently clear and precise manner; and
- real and present (i.e. not hypothetical).

Interestingly, the EDPB in its opinion on AI models do not say unequivocally that use of an AI model trained on unlawfully processed personal data is itself automatically unlawful:

"When SAs verify if and how the controller assessed the appropriateness of legitimate interest as a legal basis for the processing it conducts, the unlawfulness of the initial processing should be taken into account as part of the legitimate interest assessment, for instance by assessing the potential risks that may arise for the data subjects whose personal were unlawfully processed to develop the model"

UK: Could the Data (Use and Access) Bill recognize some AI related legitimate interest?

France: When developing an AI system, the **following interests could be considered legitimate *a priori*** (CNIL guidelines):

- carrying out scientific research;
- facilitating access by the public to certain information;
- developing new systems and functions for users of a service (to be defined with sufficient precision);
- offering a conversational agent service to assist users;
- improving a product or service to enhance its performance; or
- developing an AI system to detect fraudulent content or behaviour.

How do you do a robust LIA?

2 The interest must be "necessary"

The controller must ensure that

- **the processing is essential to meet the intended purpose** (in line with data minimisation); and
- **no less privacy-intrusive alternatives exist.**

It is important to identify the categories of data involved and in particular if special categories of data or children's data or data from other vulnerable groups will be included.

For instance, can synthetic data be an alternative?

The benefits provided by the AI system help to justify the processing. **The greater the benefits, the more likely it is that the interests will take precedence over the rights and freedoms of individuals.**

How do you do a robust LIA?

3 The interests pursued must not harm the rights and freedoms of individuals

The controller must balance:

- the **benefits** of the processing with
- its **impact** on individuals; and
- if necessary, implement **additional measures** to mitigate risks and protect individuals' rights and freedoms.

The controller should not focus the analysis on one category, but should consider all relevant:

- **individual rights and benefits**
- **stakeholders** whose rights or interests are involved (e.g., creditors, employees, civil society organisations, etc.)
- **adverse processing impacts** that may be involved to all stakeholders, their **likelihood and level of consequence** (e.g, constitutional harms, IP harms, etc.)

1. The extent and nature of the benefits expected, not only for the controller but also for third parties e.g. the users, or for the public interest.

Example: A voice recognition system that allows users to automatically transcribe their speech, such as for filling out administrative forms, can provide significant benefits by improving accessibility for individuals with disabilities.

2. The usefulness of the processing carried out for compliance with other regulations.

Example: the supplier of a VLOP or search engine who develops an AI system to better comply with the provisions of article 35.1 of the DSA on the adaptation of online content moderation processes.

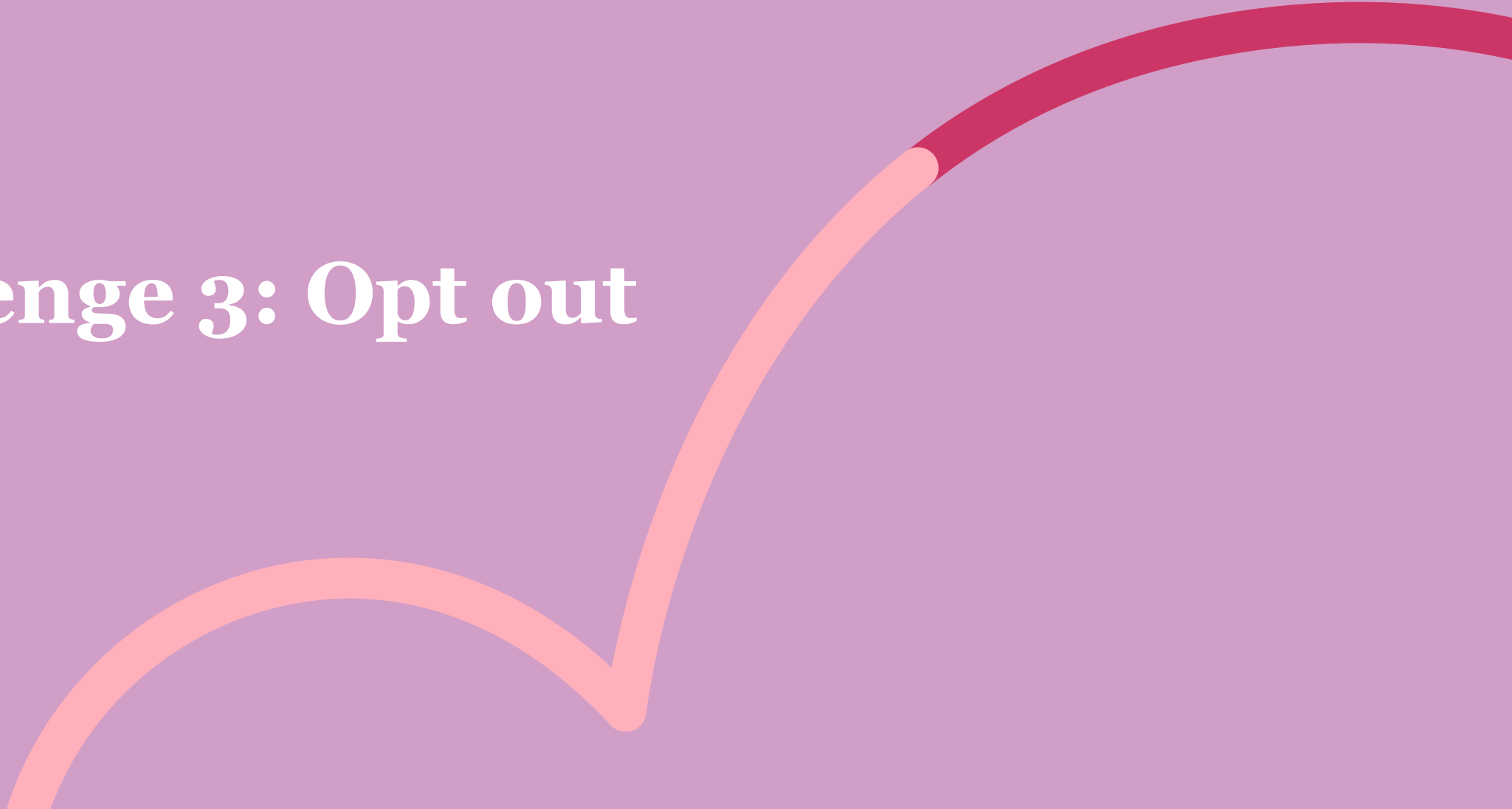
Can a starting point be found somewhere?



Yes – A draft LIA has been developed by the Information Accountability Foundation (IAF) and can be found at the following address:

https://www.informationaccountability.org/files/ugd/ffea53_dbad4aba7fe9424899b0f4f575b56c46.pdf

Challenge 3: Opt out

The background is a solid light purple color. There are two decorative wavy lines. The first is a light pink line that starts at the bottom left, curves up and to the right, and then curves down and to the right. The second is a darker red line that starts at the bottom right, curves up and to the left, and then curves down and to the left, meeting the pink line.

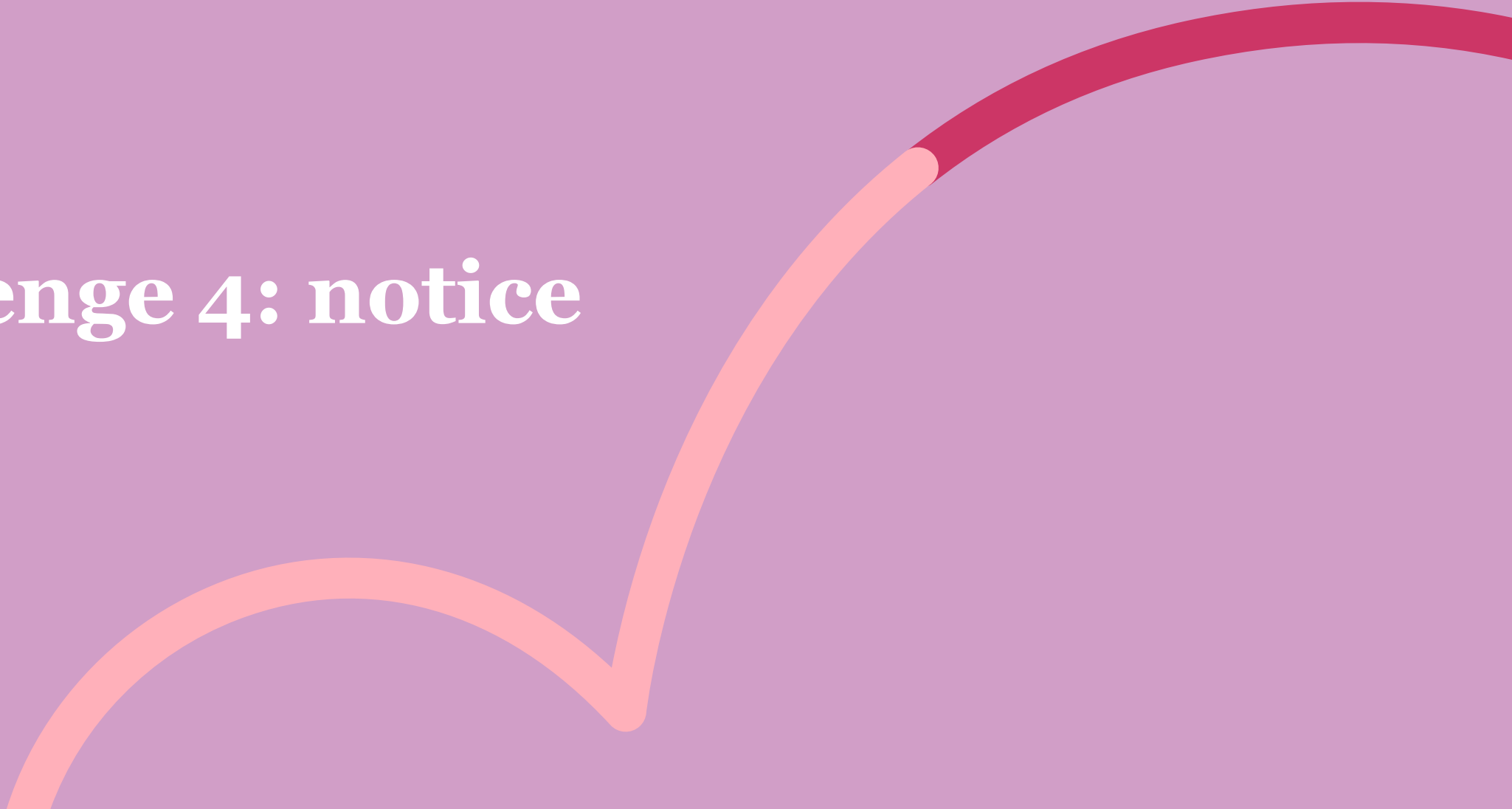
If you rely on LI, a right to object (aka opt out) must be offered unless for e.g....

- The controller can demonstrate compelling legitimate grounds for the processing, which override the interests, rights and freedoms of the individual
- This is a high bar test as reminded by the EDPB in its draft guidelines on lei:
 - The burden of proof is on the controller
 - The interest(s) must be essential. Example given by the EDPB: *"if a controller is compelled to process the personal data in order to protect its organisation or systems from serious immediate harm or from a severe penalty which would seriously affect its business"*
 - In any event, when the controller assesses whether the compelling legitimate grounds fly, the assessment of the impact of the processing must be done in light of the particular situation of the data subject. *So no in mass opt out denial.*

A right to object must therefore be provided

- On this, recently updated CNIL guidelines call out:
 - The need to make available opt out solutions ahead of web scrapping activities take place
 - If the personal data has already been ingested, the CNIL recommends:
 - Retraining the AI model without the objected personal data components (ideal scenario but difficult in practice)
 - If this can't be done, filters could be applied to ensure that the objected personal data is not used in future output

Challenge 4: notice

The background is a solid light purple. On the right side, there are two thick, wavy lines. The lower line is a light pink color, starting from the bottom left and curving upwards and to the right. The upper line is a darker red color, starting from the top right and curving downwards and to the left, meeting the pink line.

Two avenues to consider

**Direct
collection
scenario
(Art 13)**

**Indirect
collection
scenario
(Art 14)**



Thank you

twobirds.com

阿布扎比 ● 阿姆斯特丹 ● 巴黎 ● 北京 ● 布达佩斯 ● 布拉迪斯拉发 ● 布拉格 ● 布鲁塞尔 ● 迪拜 ● 东京 ● 都柏林 ● 杜塞尔多夫
● 法兰克福 ● 哥本哈根 ● 海牙 ● 汉堡 ● 赫尔辛基 ● 华沙 ● 旧金山 ● 卡萨布兰卡 ● 里昂 ● 伦敦 ● 罗马 ● 马德里
● 米兰 ● 慕尼黑 ● 上海 ● 深圳 ● 斯德哥尔摩 ● 悉尼 ● 新加坡 ● 香港

本文件中提供的技术、法律或专业事项相关信息仅供参考，不构成法律或专业意见。如有任何具体法律问题或事项，请咨询具有适当资质的律师。鸿鹄对于本文件所含该等信息概不负责，并且对于该等信息不承担任何法律责任。

本文件属于保密信息。除非另有说明，鸿鹄对本文件及其内容拥有版权。不得以任何实质形式出版、传播、摘录、再使用或复制本文件的任何部分。

鸿鹄律师事务所是一家国际律师事务所，由鸿鹄律师事务所及其附属和关联实体组成。

鸿鹄律师事务所是在英格兰和威尔士注册的一家有限合伙企业，注册号OC340318，经律师监管局 (Solicitors Regulation Authority) 授权并受其监管，编号为497264，注册办公室及主要营业地位于伦敦12 New Fetter Lane, EC4A1JP。有关鸿鹄律师事务所成员和任何经任命担任合伙人的非成员的名单及其各自的专业资质，均可在上述地址查阅。