

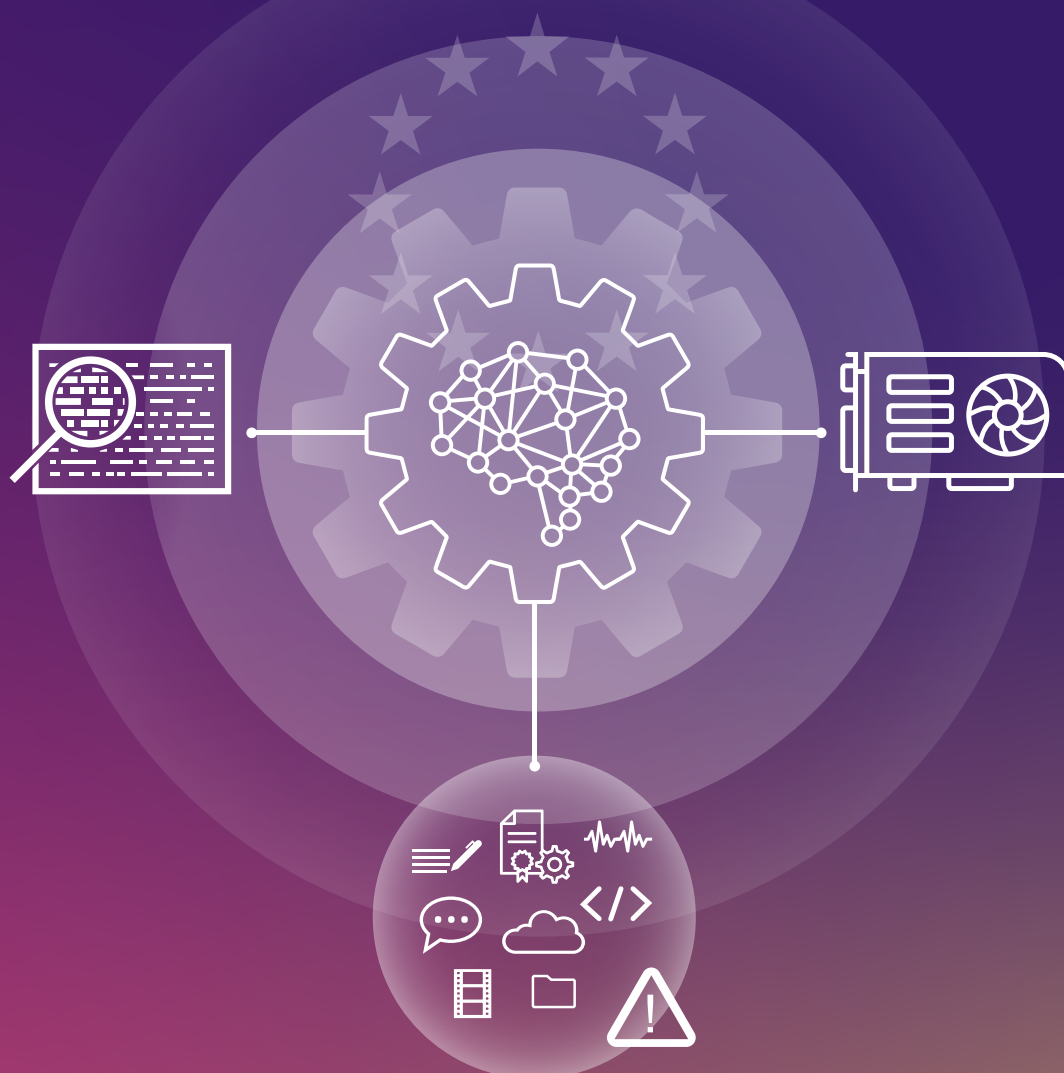
# A Framework to Categorise Modified General-Purpose AI Models as New Models Based on Behavioural Changes

Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act

Pacchiardi, L., Burden, J., Martínez-Plumed, F., Hernández-Orallo, J.

Fernández Llorca, D., Gómez, E. (editors)

**2025**



This publication is an External Study report prepared for the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

### Contact Information

Name: David Fernández Llorca

Address: European Commission, Joint Research Centre (JRC) Edificio Expo, c/Inca Garcilaso, 3, 41092 Seville - Spain

Email: david.fernandez-llorca@ec.europa.eu

### The Joint Research Centre: EU Science Hub

<https://joint-research-centre.ec.europa.eu>

JRC143257

PDF ISBN 978-92-68-31545-3 doi:10.2760/4372557 KJ-01-25-465-EN-N

Luxembourg: Publications Office of the European Union, 2025

© European Union, 2025



The reuse policy of the European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of photos or other material that is not owned by the European Union permission must be sought directly from the copyright holders.

How to cite this report: Pacchiardi, L., Burden, J., Martínez-Plumed, F., Hernández-Orallo, J. *A Framework to Categorise Modified General-Purpose AI Models as New Models Based on Behavioural Changes*, Fernández Llorca, D., Gómez, E. (editors), Publications Office of the European Union, Luxembourg, 2025, <https://data.europa.eu/doi/10.2760/4372557>, JRC143257.

**Contents**

Abstract..... 2

Acknowledgements..... 3

Note from the Editors ..... 4

Executive summary ..... 6

1 Background ..... 9

    1.1 Notion of "substantial modification" of high-risk AI systems ..... 9

    1.2 How modifications to AI models arise..... 10

2 Approach 1: Directly measuring difference in behaviour ..... 11

    2.1 Differences in capability profiles..... 12

    2.2 Differences in instance-level answers..... 13

        2.2.1 Further background information on CAPA..... 15

3 Approach 2: Proxy metrics for differences in behaviour..... 16

    3.1 What alterations have the potential to cause substantially different behaviour and performance ..... 16

        3.1.1 Finetuning methods ..... 17

    3.2 Proxy metrics..... 18

    3.3 Linking alteration metrics to downstream behavioural changes ..... 18

        3.3.1 Protocol..... 19

4 Conclusions ..... 20

References ..... 21

List of abbreviations and definitions..... 22

List of tables ..... 23

## **Abstract**

This report discusses an approach for measuring behavioural change in models following modification, which could be used as part of a broader assessment to determine when a modified General-Purpose AI (GPAI) model should be considered a new and distinct model for regulatory purposes under the EU AI Act. It presents two approaches to assess behavioural changes in altered models: (1) directly measuring differences in capability profiles or instance-level answers, and (2) using proxy metrics related to the alteration process, such as finetuning, computation, and data usage. The report highlights the challenges of establishing thresholds for determining when an altered model might be considered a new one based on behavioural change, and suggests empirical studies to validate the relationships between alteration metrics and downstream behavioural changes.

## **Acknowledgements**

The authors thank the Joint Research Centre and the AI Office team for valuable feedback and suggestions on the draft.

The editors would like to thank the JRC colleagues who have helped us develop the work resulting in this collection, the AI Office for its inputs, as well as those who have kindly agreed to review the drafts and the final external study reports.

### ***Authors***

Pacchiardi, Lorenzo

Burden, John

Martínez-Plumed, Fernando

Hernández-Orallo, José

### ***Editors***

Fernández Llorca, David

Gómez, Emilia

## Note from the Editors

The EU AI Act entered into force on 1 August 2024, with the aim of promoting innovation in and uptake of AI in the Union, while ensuring a high level of protection of health, safety and fundamental rights, including democracy and the rule of law. Chapter V of the AI Act outlines obligations for the providers of general-purpose AI (GPAI) models, which are AI models *"trained with a large amount of data using self-supervision at scale, that [display] significant generality and [are] capable of competently performing a wide range of distinct tasks [...] and that can be integrated into a variety of downstream systems or applications"*. Moreover, the chapter specifies additional obligations for the most advanced GPAI models, those that pose systemic risks, which are classified as such according to criteria established in Article 51 and Annex XIII. From 2 August 2025, the obligations for providers of GPAI models and GPAI models with systemic risk enter into application.

The European Commission's Joint Research Centre (JRC) has been providing scientific support throughout the legislative process of the AI Act since 2020. After the Council and Parliament reached a final agreement in December 2023, the JRC initiated an internal study, which included two external experts, focusing on the technical aspects of Chapter V that likely required further clarification. This study generated an internal report titled "General Purpose AI Models under the AI Act" (Hernández-Orallo et al., 2024), which provided preliminary insights into compute, generality, capabilities, and systemic risks. One of the clear conclusions of this preliminary study was that further scientific work was necessary.

Between September 2024 and June 2025, the JRC setup and managed, in close collaboration with the EU AI Office, a pool of 15 external experts with diverse expertise and backgrounds. This expert pool produced further technical scientific input on key aspects of Chapter V, through the development of methodologies for categorising AI models as GPAI models and for classifying GPAI models as GPAI models with systemic risk, to inform implementation of the EU AI Act. The experts also provided input on the recently published Commission guidelines on the scope of obligations for providers of GPAI models (European Commission, 2025), as part of the public multi-stakeholder consultation. The primary outcome of this expert pool is this **Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act**, which comprises a total of six external scientific study reports.

Although more documents may be added to the collection in the future, as of the writing of this editorial, the titles of the external reports included in the collection are:

- Training Compute Thresholds - Key Considerations for the EU AI Act
- A Framework for General-Purpose AI Model Categorisation
- A Framework to Categorise Modified General-Purpose AI Models as New Models Based on Behavioural Changes
- A Proposal to Identify High Impact Capabilities of General-Purpose AI Models
- The Role of AI Safety Benchmarks in Evaluating Systemic Risks in General-Purpose AI Model
- General-Purpose AI Model Reach as Criterion for Systemic Risk

The overall objective of this collection is twofold. On the one hand, it aims to contribute to broadening the understanding and discussion of the technical and scientific issues related to GPAI models and the identification of systemic risks. On the other hand, it seeks to provide a solid scientific basis for informing the implementation of Chapter V of the EU AI Act, which has recently entered into application. It is clear that we are dealing with complex issues, where a clear scientific consensus has yet to be established, and which require a certain degree of flexibility. Nevertheless, this is part of the

necessary effort to promote innovation and the uptake of AI, while ensuring protection for human health, safety, and fundamental rights in Europe.

These external scientific studies cover aspects regarding the presumption of having high impact capabilities based on cumulative amount of computation used for training (Article 51(2)), notification conditions (Article 52), the definition of a GPAI model (Article 3(63)), and considerations for GPAI models being classified as GPAI models with systemic risk based on capability benchmarks, safety benchmarks and reach (Article 51(1) and Annex XIII).

These studies reflect the outcome of the scientific and technical analysis of a series of external experts to the Commission. In some cases, they present a state-of-the-art review, while in others, they propose methodologies based on solid scientific evidence, while acknowledging significant uncertainty, as many of these problems still lack a widely accepted solution. The content, analysis, recommendations, and suggestions should then not be interpreted in any way as the position of the Commission, nor of the JRC editors in particular, but rather as the opinion of the authors.

## **Executive summary**

This report is motivated by the following question:

- When should a modified, adapted or fine-tuned general-purpose AI (GPAI) model be considered a distinct new model under the EU AI Act?

While different factors may need to be taken into account in answering the above question, in this report we focus on behavioural change. That is, we discuss ways to measure differences in a model's behaviour following modification, and potential proxy metrics (e.g., fine-tuning, amount of computation, data usage) for determining when the behaviour of a modified or fine-tuned model deviates sufficiently from its original version to warrant being considered a new distinct model.

This report presents a potential methodology, without delving into setting numerical thresholds and to what extent system-level components should be considered. These considerations should be determined to align our proposed methodology with legal, technical and regulatory developments. Moreover, they can be refined in the future to account for evolving technical and regulatory landscape.

AI models, being software systems, are not immutable and easily separable entities in the same way as physical products. An AI model may be modified and altered in several ways and by increasing degrees, resulting in various degrees of change to their functionalities.

There is no objective and meaningful choice for when an altered AI model must be considered a "distinct" one or not. However, whether or not an altered AI model is a new model matters for regulatory purposes since the EU AI Act includes various requirements for providers of GPAI models and of GPAI models with systemic risk (Articles 53, 55). To illustrate the purpose of our work, let's consider a provider of a GPAI model that is compliant with the obligations established in Article 53. If the provider starts developing a new version of the model based on the previous one, what alterations would cause the modified model to be considered a distinct object, subject to the obligations established in Article 53?<sup>1</sup>

When assessing whether an altered GPAI model should be considered a new GPAI model, a relevant consideration may be the extent to which its **behavioural characteristics** differ from the original one. For example, not all alterations involve modifying a model's behaviour, even if they are consequential. This could be the case if the alterations concern only changes to the way the model is implemented (e.g., different algorithms to perform the same computations more efficiently but keeping a similar pace of interaction with humans or the environment, or the use of different hardware) without affecting the model's behaviour<sup>2</sup>. Moreover, in order to actually measure the behavioural change of a model following alteration, it can be helpful for practical purposes to rely on proxy metrics which can predict the extent to which a particular alteration method will impact the downstream model's behaviour.

In light of the above considerations, we propose two approaches for assessing the extent to which an altered model's behaviour differs from the original one:

1. Directly measuring the difference in behaviour between the two models.
2. Relying on a set of proxy criteria depending on the extent and type of alteration which can be used to predict the change in behaviour (i.e. what is measured with the first approach).

The first approach requires testing the altered model extensively and comparing behavioural features to the original one, while the second approach offers a (generally) faster and cheaper approximation.

In both cases, a specific threshold must be set to determine when a model is considered a distinct one from the original model, if the decision is made to base this on behavioural differences. The question of which threshold should be set is not addressed in this report given its dependence on policy and legal, rather than technical, considerations.

For Approach 1, we must address the following question:

- What are the most relevant direct metrics for measuring behavioural changes between a model and its altered version?

For Approach 2, the following two questions should be answered:

---

<sup>1</sup>Note that, as established in recital 109 regarding commensurability and proportionality of obligations of providers of GPAI models, *"In the case of a modification or fine-tuning of a model, the obligations for providers of general-purpose AI models should be limited to that modification or fine-tuning, for example by complementing the already existing technical documentation with information on the modifications, including new training data sources, as a means to comply with the value chain obligations provided in this Regulation."*

<sup>2</sup>The possibility of such modifications relates to the Ship of Theseus paradox (Wikipedia, 2025), which asks whether a ship is the same ship after all its parts have been replaced. Similarly, one can ask whether an AI model that has had all its "parts" modified yet which preserves the same behaviour or functionalities is the same model.

- What kinds of alterations have the potential to lead to a model with substantially different behaviour and performance?
- How can such alterations be quantified by metrics that are predictive of such behavioural changes?<sup>3</sup>

For both approaches, we may also ask: how can we set quantitative thresholds such that any alteration above the threshold is considered to originate a distinct model?

Note that the modifications implemented on a GPAI model that make it a new distinct model based on behavioural change could affect its generality, and may also impact its consideration as a GPAI model (e.g. alterations that narrow the scope of tasks that the model is competent to perform). Similarly, such modifications leading to behavioural changes could alter the model's capability levels and risk profiles, potentially influencing its consideration as a GPAI model with systemic risk (e.g. because the altered model has been fine-tuned to be highly proficient in or to forget about Chemical, Biological, Radiological, or Nuclear (CBRN) knowledge). Although these aspects are of interest, and are possibly related to behavioural change measurements, we do not focus here on understanding what alterations may cause a GPAI model to cease being a GPAI model or to become, or cease to be, a GPAI model with systemic risk.

---

<sup>3</sup>If no quantitative metrics are available, proxies cannot be properly applied to measure the alteration that leads to a behavioural change.

# 1 Background

## 1.1 Notion of "substantial modification" of high-risk AI systems

The AI Act defines the concept of substantial modification of high-risk AI systems and relies on it in some of the requirements. The definition is the following (Article 3 (23)):

- *'substantial modification' means a change to an AI system after its placing on the market or putting into service which is not foreseen or planned in the initial conformity assessment carried out by the provider and as a result of which the compliance of the AI system with the requirements set out in Chapter III, Section 2 is affected or results in a modification to the intended purpose for which the AI system has been assessed.*

This would imply the following:

- A change that was foreseen and which was considered in the initial conformity assessment does not count as substantial modification.
- If the change affects compliance with the regulation, then it is to be considered a substantial modification.
- If the change modifies the intended purpose, then it is to be considered a substantial modification.

This definition does not directly translate to GPAI models, which do not have an "intended purpose". The second point broadly points to failing with complying with the regulation and is therefore uninformative for our aim. The first point may be taken in consideration: we may consider ways in which a GPAI model provider foresees some kinds and amounts of alteration to the model and puts in place compliance measures that ensure the requirements will be fulfilled even after the modification has taken place. This could be for instance used in case of models learning online (i.e., being updated during deployment based on user feedback). This is also linked to Article 43(4) on conformity assessment for high-risk AI systems: "[...] For high-risk AI systems that continue to learn after being placed on the market or put into service, changes to the high-risk AI system and its performance that have been predetermined by the provider at the moment of the initial conformity assessment and are part of the information contained in the technical documentation referred to in point 2(f) of Annex IV, shall not constitute a substantial modification". A possible technical approach to do so is developing bounds on the change in behaviour with an amount of fine-tuning and take that into account before releasing a model. However, we are unaware of currently available techniques allowing this and are sceptical that these could be developed in the near future.

See also Recital 128 on this: *"In line with the commonly established notion of substantial modification for products regulated by Union harmonisation legislation, it is appropriate that whenever a change occurs which may affect the compliance of a high-risk AI system with this Regulation (e.g. change of operating system or software architecture), or when the intended purpose of the system changes, that AI system should be considered to be a new AI system which should undergo a new conformity assessment. However, changes occurring to the algorithm and the performance of AI systems which continue to 'learn' after being placed on the market or put into service, namely automatically adapting how functions are carried out, should not constitute a substantial modification, provided that those changes have been pre-determined by the provider and assessed at the moment of the conformity assessment".*

For many complex machine learning techniques, such as deep learning, there is no guarantee in general that future adaptation to new data through learning (by non-linear gradient descent or other

similar methods) can anticipate the changes before seeing the new data. This includes finetuning and most approaches for knowledge editing and unlearning. We discuss an approach to establish relationships between alteration approaches and behavioural changes in Section 4.3.

Overall, therefore, while there is indeed a parallel between the topic of this document and the notion of substantial modifications of high-risk AI systems, without further work on this (for instance similar to what is suggested in Section 4.3), it is hard to foresee modifications to AI models and pre-emptively ensure compliance in face of those changes, due to the technical nature of the problem.

## 1.2 How modifications to AI models arise

There are two scenarios that lead to alteration of an existing model:

- **Internal modification and release:** a provider/deployer may internally modify a model (which is already present on the market) and then put the altered model on the market.
- **Post-deployment adaptiveness:** a model already deployed may change by learning while in use. In practice, the model could be either served via API by a provider, in which way the latter maintains control of the model, or being run by independent users (for instance, a model is locally run on a user's laptop and learns the user habits over time, finetuning over feedback or some other approaches). This kind of adaptiveness is what recital 12 (for AI systems) refers to: *"the adaptiveness that an AI system could exhibit after deployment refers to the self-learning capabilities, allowing the system to change while in use"*. Recital 12 refers to the definition of AI system, which includes *"that may exhibit adaptiveness after deployment"*. The European Law Institute (ELI, 2024) discusses how this could also be interpreted as, for instance, a chatbot "adapting" its answers across a multi-turn conversation considering the previous interactions; similarly, ChatGPT (an AI system) now allows users to specify what things the selected model must remember about the user itself across conversations. We contend however that these approaches should not be considered as adapting the model, as they are instead simply changing the input it sees<sup>4</sup>. Instead, a model has post-deployment adaptiveness if it is modified from the data (and possibly feedback) it is subject to, independently of whether the data comes from a single user or multiple users.

Determining whether an altered model should be considered a new one based on its behavioural change should be independent of which scenario caused the modification. However, which scenario it was may impact the suitability of the approaches presented below for assessing behavioural change.

---

<sup>4</sup>Notice that this may however be considered as adapting the AI system. This is not relevant to our focus on understanding when the underlying model has to be considered distinct.

## 2 Approach 1: Directly measuring difference in behaviour

As discussed in the introduction, the extent to which a modification changes the behaviour of a model is a relevant factor when assessing whether a modified model should be considered a new model. However, there are multiple ways to measure difference in behaviour. In this section, we present two family of direct measures, highlight their benefits and drawbacks, and discuss how they could be used to determine whether a modified model is a new model on the basis of behavioural change. The measurement methodologies that we consider are:

- Extracting capability and/or propensity profiles of the original and altered models based on the approach put forward in Section 3 of (Burden et al., 2025) to categorise AI models as GPAI models and comparing their differences using a suitable metric.
- Measuring similarity in the original and altered models' instance-level answers on benchmarks<sup>5</sup> (for instance, using Chance Adjusted Probability Agreement (Goel et al., 2025)). Here, we would directly compare how differently two models answer to individual instances.

These two measures are mostly orthogonal: the first measure focuses on capability profiles and ignores therefore the fact that different models with the same capability profiles may answer differently on individual instances (as long as the ratio of correct answers is similar at each "difficulty level", as that is where the capabilities are extracted from). For this same reason, the second measure could yield high values even if the capability profiles of the altered model is preserved. Conversely, a relatively small difference obtained by the second measure may correspond to a substantial change in the former if, for instance, the altered model answers identically to the original one for all instances up to a certain level of difficulty but improves drastically on the next difficulty level, thus leading to an increase in the capability estimate obtained with the former approach. Overall, therefore, an increase in instance-level difference is not monotonically linked to an increase in capability difference, and vice-versa.

Additionally, while instance-level similarity metrics have been shown to be informative of downstream effects when two models interact (see Section 3 in (Goel et al., 2025)), the capability profiles obtained with the approach in Section 3 in (Burden et al., 2025) have been shown to be predictive of performance in Out-Of-Distribution scenarios.

Finally, the second measure depends more strongly on the considered choice of benchmarks (as the responses of the two models may differ more on particular benchmarks), while the former does so less, as long as the annotated demand levels are predictive of difficulty. The former measure also relies on the same set of evaluations involved in the categorisation of AI models as GPAI models presented in (Burden et al., 2025), offering the ability to leverage the same results for two purposes.

Overall, the choice between what kind of family measures to rely upon for determining the extent to which a modified model differs in behaviour from the original model, and in turn when an altered model might be considered a new model based on these differences, is a normative choice:

- Capability-profile difference measures are suitable if we want an altered model to be considered "distinct" in case its behaviour changes broadly on a large set of tasks, but not if the alteration only caused the model to become better at a few specific tasks. If there is no significant change on broad sets of tasks, the altered model would not be considered a new model.
- Instance-level difference measures focus on narrower changes and are therefore suitable if narrow changes of this kind are deemed important for determining whether an altered model

---

<sup>5</sup>A model's instance-level answer refers to the model's output for a particular example, instantiation, item, or question corresponding to a benchmark.

should be considered a new model. Moreover, their higher sensitivity to the used benchmark implies that the method can be adapted to look for changes in specific domains. If there is no significant change in this domain, we would determine that the modified model is not a new model, at least for that specific domain.

On the basis that GPAI models are characterised by their ability to tackle a large range of tasks—Article 3(63) of the EU AI Act (European Union, 2024)—, we believe using capability-profile difference measures may be a more appropriate method of measuring behavioural change in modified models. Relying on instance-level metrics would necessarily require either considering a narrow set of tasks or having a huge selection of benchmarks; the former would not be representative of overall behaviour change, while the second would be impractical. By contrast, capability profiles are better predictive of broad changes in downstream performance. Moreover, it is convenient that capability profiles also underpin the proposed procedure for categorising AI models as GPAI models (Burden et al., 2025), meaning that the same techniques could be used both for this task, and for assessing whether a modified GPAI model is a new GPAI model.

A possible intermediate option would be to extend the capability-profile with a propensity profile and other high-level dimensions that characterise the behaviour of the model, beyond its capabilities. For instance, a system may stay the same in terms of capabilities, but its propensities (and hence risk) may have changed. This affects the classification of GPAI models as GPAI models with systemic risk more than the categorisation of AI models as GPAI models, but it would be generally better than the second approach above. However, we leave exploring such an approach for future work.

Below, we discuss in detail how both measures (capability profile differences and instance-level difference) could be implemented. We also point out that there is a range of options between these two approaches and in the ways in which the profiles are derived or compared, and that the behaviours can be compared at a more granular level. Note that thresholds must be established, since the same model with some stochasticity will differ in specific inputs, but minor changes in expectation may be considered insufficient to consider them different models.

## 2.1 Differences in capability profiles

This method involves obtaining the capability levels for the four cognitive domains as discussed in Section 3 in (Burden et al., 2025)<sup>6</sup>. As this is part of the procedure proposed approach for the categorisation of AI models as GPAI models, this is an efficient method since it relies on evaluations that must already be carried out for this other goal.

Once the capability levels for the original and altered models are obtained, these can be grouped into vectors  $C_o$  and  $C_a$  and compared with the use of a metric:

$$d(C_o, C_a) \tag{1}$$

Once could then set a threshold such that, if this "distance" is larger than the set threshold, the altered model would be considered a new and distinct model (or considered to meet one criterion amongst others for being considered a new and distinct model).

Multiple possible metrics could be used:

- The **L1 distance (Manhattan)**  $d_1(C_o, C_a) = \sum_{i=1}^n |C_{o,i} - C_{a,i}|$  (where  $n$  is the number of considered capabilities) represents the sum of the absolute difference between the values for

---

<sup>6</sup>While one could consider the full list of domains in (Zhou et al., 2025) or (Tolan et al., 2021), we believe using the same four domains as in (Burden et al., 2025) makes the approach more applicable at the cost of a minor decrease in sensitivity.

each capability. This distance has an "additive" property: if two subsequent alterations from  $C_o \rightarrow C_a^1 \rightarrow C_a^2$  either affect different capabilities or alter each of the affected capability in the same direction (e.g., capability 1 is increased by both alterations and capability 2 is decreased by both), then the difference between the original and final model is identical to the difference between the original and intermediate model plus the one between the intermediate and final model:  $d_1(C_o, C_a^2) = d_1(C_o, C_a^1) + d_1(C_a^1, C_a^2)$ . In contrast, other metrics such as **L2 (Euclidean)** or **L $\infty$  (Chebyshev)** do not exhibit this additive property, making L1 uniquely suitable to quantify incremental or sequential capability modifications.

- The **L $\infty$  (Chebyshev)**  $d_\infty(C_o, C_a) = \max_{\{i=1..n\}} |C_{o,i} - C_{a,i}|$  (where  $n$  is the number of considered capabilities) represents the maximum absolute difference across all capabilities. This metric captures the largest single deviation between the original and altered models, regardless of how small or large the other changes are. Unlike the L1 distance, the L $\infty$  norm does not have an additive property. Instead, the Chebyshev distance reflects the dominant change among all dimensions, which makes it useful when the focus is on the most significant individual change rather than the total cumulative effect. While this can be valuable for identifying outliers or ensuring no single capability has changed beyond a threshold, it is less suitable for tracking the incremental or additive impact of multiple sequential modifications.

Given the considerations above, we propose using the L1 norm as this considers changes in all capabilities and has the additive property.

Of course, a threshold should be determined. We refrain from doing so here as this is mostly a policy choice and may require to be informed from data (to which we do not have access) about the number of altered models that would be considered as distinct with each threshold choice.

## 2.2 Differences in instance-level answers

Another method for measuring model similarity is that of testing the original and altered model on a (set of) multiple-choice benchmarks, recording the probabilities the models assign to each answer (not only the answer they produce or whether they are correct or incorrect) on each benchmark instance (or sample), and computing a metric of similarity. In particular, a possible strategy involves computing CAPA (Chance Adjusted Probability Agreement), introduced in a recent work (Goel et al., 2025) to compare pairs of models. CAPA extends Cohen's kappa by considering output probabilities and considering the probability assigned to all options of each multiple-choice question (e.g., the labels A, B or C)<sup>7</sup>. CAPA normalises the observed raw probabilistic agreement  $c_{obs}^p$  and the "expected" one  $c_{exp}^p$  obtained if the models were fully independent but maintain the observed average probability assigned to the correct option (and uniformly split the remaining probability across the other options). In particular, CAPA is defined as:

$$\kappa_p = \frac{c_{obs}^p - c_{exp}^p}{1 - c_{exp}^p} \quad (2)$$

where:

$$c_{obs}^p = \frac{1}{|D|} \sum_{x \in D} \sum_{o_i \in O(x)} p_1(o_i) \cdot p_2(o_i) \quad (3)$$

and:

<sup>7</sup>A Python package to compute the similarity is also available: <https://pypi.org/project/lm-sim/>

$$c_{exp}^p = \underbrace{\overline{p_1} \cdot \overline{p_2}}_{\text{chance agreement on correct option}} + \underbrace{(1 - \overline{p_1}) \cdot (1 - \overline{p_2}) \cdot \frac{1}{|D|} \sum_{x \in D} \frac{1}{|O(x) - 1|}}_{\text{chance agreement on incorrect option}} \quad (4)$$

with  $x$  denoting the data sample,  $O(x)$  denoting the options for sample  $x$ ,  $p_1(o_i)$  and  $p_2(o_i)$  denote the output probabilities for models 1 and 2 respectively,  $|D|$  is the number of data points,  $|O(x)|$  is the number of options for sample  $x$ , and  $\overline{p_j}$  is the average probability assigned to the correct options by model  $j$  (notice that, if a model is perfectly calibrated,  $\overline{p_j}$  corresponds to the accuracy). CAPA has the following properties:

- It is between  $-1$  and  $1$ , with  $-1$  indicating perfect disagreement,  $0$  indicating the agreement obtained by two independent models, and  $1$  indicating perfect agreement (both models always assign probability  $1$  to the identical choice for each sample). As  $\kappa_p$  increases, it means models make more similar choices, their errors become more correlated, and they are functionally less different.
- It adjusts for accuracy by rescaling the raw similarity measure; this ensures that, if a pair of high-performing models is compared to a pair of models achieving performance close to  $1/|O(x)|$  (i.e., close to random performance), the high-performing models are not scored as more similar to each other only because they agree more by chance. Moreover, this allows to compare on a similar scale model similarity over different benchmarks with different baseline accuracies.
- It considers model probabilities (if these are not available, the "discrete" CAPA assigning probability  $1$  to the model's output can be used).
- It distinguishes between different mistakes, thus avoiding issues with previous metrics that considered as agreement models providing different wrong predictions to the same question.

In (Goel et al., 2025) the authors find that increasing CAPA similarity positively biases a model used to score the output of a second one and reduces the gain obtained by training a model on annotations from a second model. While these trends were partly present when adopting other similarity metrics, CAPA showed the strongest predictive power thanks to incorporating the most information (such as considering probabilities and distinguishing different mistakes).

These findings indicate how instance-level similarity on benchmark instances is informative of aggregate behavioural changes in various use cases and makes it a compelling approach to determine when an altered model should be considered a distinct model for regulatory purposes. Therefore, in the specific case of our interest, we can use CAPA to measure the similarity between the original and altered model and we can consider the altered model as a distinct one whenever the similarity is below a fixed threshold.

Moreover, the obtained values strongly depend on the chosen benchmarks<sup>8</sup>. The benchmarks can either be chosen to be large generalist benchmarks, aiming to represent the large variety of tasks that GPAI models can be applied to (for instance, in (Goel et al., 2025) the authors used MMLU and MMLU-Pro and Big-Bench Hard), and/or a set of small benchmarks focusing on specific tasks that are considered as particularly relevant for the determination of whether an altered model should be considered a new model.

When computing the probability of the various options for a benchmark, we recommend extracting the probability the AI model assigns to the various option labels (e.g., "A", "B", or "1", "2") rather than to the actual answer ("Paris" or "Rome") to avoid the issue of the probability mass being split into

<sup>8</sup>See (Eriksson et al., 2025) for an interdisciplinary review of current issues with AI benchmarks.

multiple tokens. Moreover, we suggest normalising the probabilities by the total probability assigned by the AI model to all the options (e.g.,  $\hat{p}(A) = p(A)/(p(A) + p(B) + p(C))$ , if 3 options are present; this corresponds to a conditional probability) to exclude the probability mass assigned to tokens not representing the option.

As for the capability profile difference approach, using CAPA to determine if an altered model should be considered as distinct requires deciding on a threshold. As before, defining this threshold is a policy decision that will benefit from experimental data.

### 2.2.1 Further background information on CAPA

CAPA is related to Cohen’s kappa (to which it corresponds when two options are present and discrete CAPA is used) and generalises the error consistency metric  $\kappa = \frac{c_{obs} - c_{exp}}{1 - c_{exp}}$ , with  $c_{obs}$  denoting the fraction of samples on which both models are correct or wrong at the same time and  $c_{exp} = acc_1 \cdot acc_2 + (1 - acc_1) \cdot (1 - acc_2)$ , by taking into account model probabilities and distinguishing different mistakes. Chiefly, CAPA differs from Cohen’s kappa as the latter assumes fixed categories throughout all samples and computes marginal probability distributions per category. However, multiple-choice benchmarks do not generally have an inherent category ‘a’ or ‘b’, (as options can be permuted), so we cannot compute such marginal probability distributions. Moreover, Cohen’s kappa does not consider any class as particularly relevant (as it is the case with CAPA for correct answers). The authors in (Goel et al., 2025) propose an extension of CAPA for classification settings (i.e., when the classes are fixed throughout samples) in Appendix A.3.

The vanilla definition of CAPA can only be applied to multiple-choice benchmarks and assumes the model can produce probabilities for the considered choices. As mentioned above, the discrete CAPA assigns probability 1 to the model’s output (and 0 to all other choices) and can therefore be used when probabilities are unavailable; however, in (Goel et al., 2025) the authors find that this is less predictive of changes in downstream behaviour. When the considered benchmark is not multiple-choice (for instance, it considers exact match), they suggest computing the discrete CAPA and assume there is an infinite number of incorrect choices, so the probability of matching on incorrect choices is 0. However, the use of this extension was not tested. Further details are given in Appendix A in (Goel et al., 2025).

### 3 Approach 2: Proxy metrics for differences in behaviour

In this section, we first discuss what alterations have the potential to cause substantially different behaviour; then, we analyse what proxy metrics can be used to measure the extent of those alterations. We then discuss whether those metrics are likely to be well-predictive of the downstream change in behaviour (quantified by the different capability profiles, as discussed in Section 2) and put forward considerations for how thresholds for these proxy metrics could be set to reflect thresholds set for the direct metrics.

#### 3.1 What alterations have the potential to cause substantially different behaviour and performance

The following alterations cannot lead to a model with substantially different behaviour and performance, as they only affect structural and algorithmic properties of the model:

- Changing format by which the model is stored in digital memory.
- Changing architecture and hardware of computer chips without impacting the output.
- Changing the algorithms to perform computations to produce outputs from the model (e.g., batch size, vectorial optimisations), as long as the results are identical.

Similarly, the following system-level techniques do not directly modify the underlying model (although they may lead to better performance on some tasks):

- Better prompting approaches (e.g., chain of thought, few-shot examples).
- Scaffolding and tool use: these structure the model's reasoning, potentially involving multiple copies of a model, and change a model's affordances.
- Different algorithms for sampling from the model: E.g.: structured outputs rather than normal sampling, multiple sampling with solution choice, or model steering from activations.

By contrast, the following model alterations impact the behaviour and performance of the model:

- **Finetuning**, which involves modifying all or a subset of the weights of the original model using some training algorithm. Several finetuning algorithms and approaches exist, including via supervised objectives, RL, **distillation** of a teacher model into a student model, or other approaches (this can also include teaching a model to use tools and then give the model access to those tools, see (Davidson et al., 2023)).
- **Techniques to remove information from models**, such as model unlearning (Yao et al., 2024).
- **Techniques to simplify models**, such as pruning (Ma et al., 2023) and quantisation (i.e., reducing the numerical precision by which model weights are stored).

The list above is not exhaustive, there may be other current or future model alteration techniques that should be considered.

Notice that the last two alterations most likely degrade model performance, and may even affect the model's ability to perform a wide range of tasks.

Below, we give an overview of the various fine-tuning approaches, as this is one of the most popular approaches.

### 3.1.1 Finetuning methods

Finetuning methods can be categorised according to the training objective (and data) they consider and the algorithm they use to modify the model's weight. Here we give a quick overview of the most common training objectives and training algorithms, noting that this is a vast and quickly expanding field, so this cannot be exhaustive.

Training objectives define why and what signals or data are employed to refine the model. Popular approaches include:

1. **Supervised Fine-Tuning (SFT)** involves training the model using labelled datasets, commonly in the form of input-output pairs, such as instruction-response examples. The primary goal here is to minimize a supervised loss function, like cross-entropy, ensuring the model closely aligns with the provided labels or instructions.
2. **Reinforcement Learning Fine-Tuning** employs a reward-based signal instead of explicit labelled data. A prominent example is Reinforcement Learning from Human Feedback (RLHF), where human judgments act as the reward signals guiding model improvements.
3. **Distillation-Based Fine-Tuning** is characterised by refining a less capable "student" model to replicate the behaviour of a more capable "teacher" model. The student model typically learns by matching either the logits or the internal representations generated by the teacher, leveraging the teacher's superior performance while reducing computational demands.
4. **Self-Supervised** or **Unsupervised Fine-Tuning** continues model training using unlabelled textual data through tasks like next-token prediction. This method requires no external labelled datasets or explicit reward signals, relying solely on inherent structure and patterns within unlabelled data.

For each of those training objectives, different algorithms to update model weights can be used. Popular approaches include:

1. **Full Parameter Fine-Tuning** updates every parameter in the model through standard gradient descent. This straightforward approach ensures comprehensive adaptation but can be computationally demanding, particularly with large-scale models.
2. **Low Rank Adaptation (LoRA)-Based Methods** improve computational efficiency by inserting trainable, low-rank adaptation matrices into selected layers, with only these newly introduced matrices undergoing updates. A variant, QLoRA (Quantised LoRA), utilises 4-bit quantisation to further reduce GPU memory usage while maintaining high efficiency in parameter updates.
3. **Adapter-Based Methods** incorporate compact bottleneck modules or adapters into each transformer block, updating only these adapter parameters while freezing the underlying model's weights. **BitFit**, an even more streamlined alternative, restricts parameter updates solely to bias terms, maximising parameter efficiency albeit at some potential cost to flexibility.
4. **Prompt and Prefix Tuning** represent lightweight parameter-update strategies, where additional learned embeddings or "prefixes" are introduced into the model. **Prefix Tuning** inserts trainable prefix vectors at specific layers, leaving the base model parameters unchanged, whereas **Prompt Tuning** restricts learning to embedding vectors prepended or appended only to the input layer.

### 3.2 Proxy metrics

For each of the alterations listed above, different metrics could be used to quantify the amount of alteration, including:

- Monetary cost of modification.
- Amount of compute.
- Amount and variety of data used, if applicable (not all alteration methods are based on data).
- "Strength" of the alteration applied.

These could all be absolute or relative with respect to some properties of the starting model, e.g., size or initial training cost.

Some of these metrics apply to all alterations listed above (such as the amount of compute used), while some are specific to a subset (e.g., model pruning and quantization does not involve data). Table 1 shows what metrics apply to each alteration type (tick), which do not apply (cross) and which may apply depending on the specific algorithm, used for that alteration technique (question mark).

**Table 1:** Metrics that apply to each alteration type (✓), which do not apply (✗) and which may apply depending on the specific algorithm used for that alteration technique (?).

	<b>Monetary cost</b>	<b>Compute cost</b>	<b>Data used</b>	<b>Variety of data used</b>	<b>Other metrics of alteration strength</b>
<b>Finetuning</b>	✓	✓	✓	✓	✓ (update size, number of parameters changed)
<b>Distillation</b>	✓	✓	✓	✓	✓ (update size, number of parameters changed, teacher model features)
<b>Unlearning</b>	✓	✓	✓	✓	✓ (strength of unlearning)
<b>Pruning</b>	✓	✓	?	?	✓ (strength of pruning)
<b>Quantisation</b>	✗	✗	✗	✗	✓ (strength of quantisation)

*Source: Own elaboration.*

The amount of data used can be measured in terms of number of tokens or training samples. Instead, the variety of data could be measured either by simple work-level metrics, or by considering, for instance, the distribution of demand levels obtained by the approach presented in Section 3 in (Burden et al., 2025).

### 3.3 Linking alteration metrics to downstream behavioural changes

In the section above, we listed the metrics that apply to each alteration type. In this section, we instead look at connecting such metrics with the downstream behavioural changes (as measured by the difference in capability profiles).

However, different methods for a particular alteration type (e.g., the different fine-tuning approaches listed above) may be more or less "efficient" in terms of a particular metric (for instance, Low-Rank

adaptation is less costly than full fine-tuning). This makes it hard to directly link the metrics above to the downstream behavioural change without considering the particular algorithm used.

It is also essential to stress how there is intrinsic asymmetry in alteration direction: adding new skills or knowledge to an AI model is much harder than removing it; therefore, the intended direction of the considered alteration approach should be considered in determining a suitable relationship with the downstream behavioural change.

The effect of alterations on downstream model behaviour must, therefore, be measured in empirical studies. Below, we outline an experimental protocol to empirically determine this. While it is impossible to apply this experimental protocol to all kinds and instantiations of model alteration, we believe that most altered models are actually obtained with a small subset of alteration methods (popular methods seem to be supervised and RL fine-tuning with LoRA, distillation approaches and quantization approaches). Therefore, the regulator could concentrate on empirically studying these more popular approaches.

### 3.3.1 Protocol

For a specific alteration algorithm (e.g., supervised finetuning with LoRA or quantisation), the following protocol allows to establish a link to the downstream behavioural changes of the model:

- a. Determine the set of metrics/indicators that characterise this alteration.
- b. Design a set of experiments varying them systematically.
- c. Consider a set of models and apply the experiments to those models.
- d. Measure the difference in capability profiles between the original and altered models over that set of experiments.
- e. If there is enough predictive power from one (or a combination of) alteration metrics to the downstream change in capability profiles, set a threshold based on the capability profile threshold.

For instance, a possible experimental setup to study how quantisation affects a model's profile would involve applying varying levels of post-training quantisation (e.g., 8-bit, 6-bit, 4-bit) to a set of pretrained language models such as GPT-2, OPT-1.3B, and LLaMA-7B. The alteration metrics would include bit-width and quantisation scheme (e.g., symmetric vs. asymmetric). These models would then be evaluated on the ADeLe battery (Section 3 in (Burden et al., 2025)) to determine their capability profile and the difference of this from the capability profile of the altered model would be computed. If this difference can be reliably predicted from quantisation metrics, thresholds could be set to ensure acceptable behavioural change limits.

## 4 Conclusions

The question underpinning this report is that of when an alteration of a GPAI model is sufficient to consider the altered model a distinct one from the perspective of the EU AI Act. Our proposal is based on the assumption that, when assessing whether a modified GPAI model should be considered a new GPAI model, a relevant consideration may be the extent to which its behavioural characteristics differ from the original one.

As a consequence, the approach we proposed involves measuring the behavioural change using either the difference in the models' capability profiles (where the capabilities are constructs broadly describing a model's performance on a range of tasks of a given difficulty) or the change in instance-level answers to the elements of specific benchmarks. We believe the former is better suited to answer the question, as it captures broader changes and depends less on the benchmarks used for the evaluation.

We also investigated a simpler approach based on proxy metrics relying on quantities describing the alteration process, but we found that establishing these relationships requires empirical validation, as it cannot be done without evidence. Therefore, we detailed an experimental protocol to conduct such studies.

All methods above require a normative choice of threshold above which a model might be considered a distinct model, or considered to meet one criterion amongst others for determining whether it should be a distinct model. While there is no absolute "right" answer, any decision will depend on policy and legal considerations.

In general, determining if a model's behaviour has changed significantly through an alteration may be moderately burdensome, unless strong proxy metrics are established through the protocol we proposed. Therefore, we encourage future empirical work investigating proxy metrics for the most common kinds of model alterations.

## References

- Burden, J., Pacchiardi, L., Martínez-Plumed, F. and Hernández-Orallo, J., 'A framework for general-purpose ai model categorisation'. In 'Collection of External Scientific Studies on General-Purpose AI Models under the EU AI Act', , edited by D. Fernández Llorca and E. Gómez, European Commission, Joint Research Centre, JRC143256, 2025.
- Davidson, T., Denain, J.-S., Villalobos, P. and Bas, G., 'Ai capabilities can be significantly improved without expensive retraining', [arXiv preprint arXiv:2312.07413](https://arxiv.org/abs/2312.07413), 2023.
- ELI, 'Commission Guidelines on the Application of the Definition of an AI System and the Prohibited AI Practices Established in the AI Act - Response of the European Law Institute'. <https://www.europeanlawinstitute.eu/>, 2024. Accessed: 2025-07-14.
- Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gómez, E. and Fernández-Llorca, D., 'Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation', [Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society \(AIES-25\)](https://proceedings.aai.acm.org/), 2025.
- European Commission, 'Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act)'. <https://ec.europa.eu/newsroom/dae/redirection/document/118340>, 2025.
- European Union, 'Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence'. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>, 2024.
- Goel, S., Struber, J., Auzina, I. A., Chandra, K. K., Kumaraguru, P., Kiela, D., Prabhu, A., Bethge, M. and Geiping, J., 'Great models think alike and this undermines ai oversight', [arXiv preprint arXiv:2502.04313](https://arxiv.org/abs/2502.04313), 2025.
- Hernández-Orallo, J., Sevilla, J., Gómez, E. and Fernández-Llorca, D., 'General-Purpose AI Models in the AI Act: Capabilities, Generality, Systemic Risks and Compute', [European Commission, Joint Research Centre, JRC139341](https://www.europeanlawinstitute.eu/), 2024.
- Ma, X., Fang, G. and Wang, X., 'Llm-pruner: On the structural pruning of large language models', [Advances in neural information processing systems](https://arxiv.org/abs/2305.13011), Vol. 36, 2023, pp. 21702–21720.
- Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J. and Gómez, E., 'Measuring the occupational impact of ai: tasks, cognitive abilities and ai benchmarks', [Journal of Artificial Intelligence Research](https://arxiv.org/abs/2105.08101), Vol. 71, 2021, pp. 191–236.
- Wikipedia, 'Ship of Theseus'. [https://en.wikipedia.org/wiki/Ship\\_of\\_Theseus](https://en.wikipedia.org/wiki/Ship_of_Theseus), 2025.
- Yao, Y., Xu, X. and Liu, Y., 'Large language model unlearning', [Advances in Neural Information Processing Systems](https://arxiv.org/abs/2405.14202), Vol. 37, 2024, pp. 105425–105475.
- Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y., Sun, L., Prunty, J. E. et al., 'General scales unlock ai evaluation with explanatory and predictive power', [arXiv preprint arXiv:2503.06378](https://arxiv.org/abs/2503.06378), 2025.

## **List of abbreviations and definitions**

**AI** Artificial Intelligence

**GPAI** General-Purpose Artificial Intelligence

**LLM** Large Language Model

**LoRA** Low Rank Adaptation

**QLoRA** Quantised Low Rank Adaptation

**RL** Reinforcement Learning

**RLHF** Reinforcement Learning from Human Feedback

**SR** Systemic Risks

**SFT** Supervised Fine-Tuning

**List of tables**

**Table 1.** Metrics that apply to each alteration type (✓), which do not apply (✗) and which may apply depending on the specific algorithm used for that alteration technique (?). . . . . 18

## Getting in touch with the EU

### In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online ([european-union.europa.eu/contact-eu/meet-us\\_en](https://european-union.europa.eu/contact-eu/meet-us_en)).

### On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: [european-union.europa.eu/contact-eu/write-us\\_en](https://european-union.europa.eu/contact-eu/write-us_en).

## Finding information about the EU

### Online

Information about the European Union in all the official languages of the EU is available on the Europa website ([european-union.europa.eu](https://european-union.europa.eu)).

### EU publications

You can view or order EU publications at [op.europa.eu/en/publications](https://op.europa.eu/en/publications). Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre ([european-union.europa.eu/contact-eu/meet-us\\_en](https://european-union.europa.eu/contact-eu/meet-us_en)).

### EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex ([eur-lex.europa.eu](https://eur-lex.europa.eu)).

### EU open data

The portal [data.europa.eu](https://data.europa.eu) provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

# Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society



Scan the QR code to visit:

**[The Joint Research Centre: EU Science Hub](https://joint-research-centre.ec.europa.eu)**

<https://joint-research-centre.ec.europa.eu>



Publications Office  
of the European Union