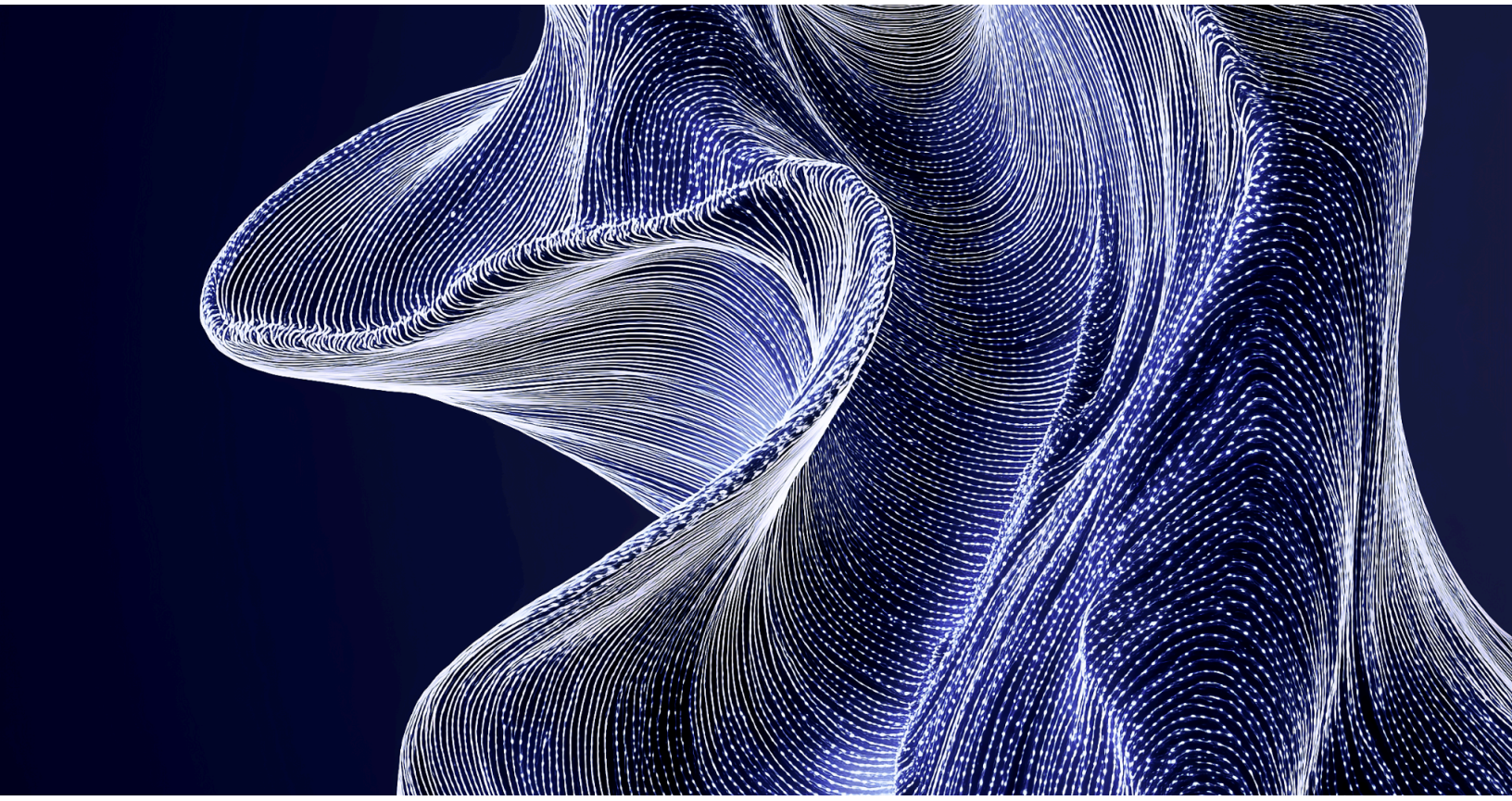
POLICY BRIEF | MARCH 2026

Requirements for Model Specifications in the EU GPAI Code of Practice

Alan Chan



GovAI

Executive Summary

The EU GPAI Code of Practice commits Signatories to providing a description of intended model behavior (a "model spec") in their Model Reports (Measure 7.1, point (4)), including:

1. The principles that the model is intended to follow,
2. How the model is intended to prioritize different kinds of principles and instructions (for example between user instructions and the system prompt),
3. Topics on which the model is intended to refuse instructions, and
4. The system prompt.

These requirements raise several questions, such as:

- To what level of detail must Signatories specify principles?
- Since a single model can be deployed with different system prompts, which one(s) must be provided?

To answer these questions, I argue that the purpose of specs is to help the AI Office (AIO) assess whether a Signatory's systemic risk mitigations are appropriate, and that the Code therefore requires specs to:

1. Include all intended behavior with significant relevance to the systemic risks posed by the model (e.g. intended limits on autonomy to mitigate loss of control risk).
2. In specifying principles, include all details that could have a significant impact on the systemic risks posed by the model (e.g. clarifying whether a restriction on bomb-making assistance still permits general information about bombs).
3. State how the model is intended to manage significant potential conflicts between principles and instructions (e.g. "Be honest" vs. "Withhold dangerous information").
4. Include system prompts for all of the Signatory's AI systems that integrate the model (e.g. web interface, mobile apps, API).

I collate these criteria into a rubric for assessing specs (see Table 1). These criteria could help Signatories adhere to the Code, the AIO interpret it, and non-Signatory frontier AI companies comply with the EU AI Act since the Code will be a benchmark for compliance.

Criteria	Guidance
Inclusion of significant intended behaviors: Does the spec include all intended principles, prioritization, and refusal topics that could have a significant impact on the systemic risks posed by the model? (Measure 7.1, point (4)(a)-(c))	Evidence from the following sources could indicate whether intended behaviors were omitted from the spec: • Public advertisements from the Signatory about the model's capabilities • The Safety and Security Framework (e.g. intentions to run particular evaluations) • The Model Report (e.g. evaluations)
Thorough specification of principles: Do the principles include all details that could have a significant impact on the systemic risk posed by the model? (Measure 7.1, point (4)(a))	For example, by itself the principle "Do not assist with the construction of CBRN weapons" is ambiguous about whether the model is allowed to provide general factual information about CBRN weapons. Principles relating to the specified systemic risks in Appendix 1.4 (e.g. loss of control risk) potentially deserve a greater level of detail. Accompanying descriptions, examples, or exceptions can provide more detail.
Management of significant conflicts: Does the spec state how the model is is intended to manage significant potential conflicts between principles and instructions? (Measure 7.1, point (4)(b))	Examples of significant potential conflicts are: • "Be honest" vs. "don't offer instructions on how to build a bomb" • User instructions vs. principles for preventing risky activities • User instructions vs. external data sources (e.g. prompt injections)
System prompts for all AI systems: Does the spec include system prompts for all of the Signatory's AI systems that integrate the model? (Measure 7.1, point (4)(d))	Has the Signatory neglected to include a system prompt for any of their AI systems? Is there reason to suspect that any of the system prompts are not in use? For example, are they largely inconsistent with system prompts that third parties have extracted and released?

Table 1: Proposed Criteria for Assessing Model Specs under the GPAI Code

Introduction

Model developers are increasingly describing how they intend their models to behave. OpenAI has its Model Spec and Anthropic has recently updated Claude's Constitution.

The EU GPAI Code of Practice ("Code") formalizes this practice: Signatories of the Code commit to providing a specification of how they intend their models to operate ("model specification" or "spec") in their Model Reports. This commitment applies to general-purpose AI models with systemic risk (GPAISR).[1] In particular, pursuant to Measure 7.1, point (4), specs must include:

1. The principles that the model is intended to follow,
2. How the model is intended to prioritize different kinds of principles and instructions,
3. Topics on which the model is intended to refuse instructions ("refusal topics"), and
4. The system prompt.

These requirements raise several questions:

- To what level of detail must Signatories specify principles?
- What does it mean to prioritize different kinds of principles and instructions?
- Since a single model can be deployed with different system prompts, which one(s) must be provided?

Answering these questions could help Signatories comply with the Code, the EU AI Office (AIO) interpret it, and non-Signatory frontier AI companies comply with the EU AI Act, since the Code of Practice will be used as a benchmark for compliance even for non-Signatories.[2]

This policy brief aims to answer these questions. After explaining key terms, I examine the purpose of Measure 7.1, point (4) and use it to interpret how the Code answers these questions.

Explaining Key Terms

Principles (Measure 7.1, point (4)(a)) are rules for model behavior, such as:

- "Don't be sycophantic" (OpenAI Model Spec)

[1] A sufficient condition for a model to be considered a GPAISR is if it was trained with more than 10^{25} FLOP. See Art. 51, EU AI Act and Section 2.3 of the Commission Guidelines for further details.
[2] From Section 5.1 of the Commission Guidelines: "Providers of general-purpose AI models that do not adhere to a code of practice that is assessed as adequate [...] are expected to explain how the measures they implement ensure compliance with their obligations under the AI Act, for instance by carrying out a gap analysis that compares the measures they have implemented with the measures set out by a code of practice that is assessed as adequate."

- "Claude only sincerely asserts things it believes to be true" (Claude's Constitution)

Instructions (Measure 7.1, point (4)(b)) are directions given to the model, which can come from sources such as users, downstream developers, or external data sources (e.g. websites accessed by the model).

Common examples of topics on which the model is intended to refuse instructions ("**refusal topics**") (Measure 7.1, point (4)(c)) include sexual content involving minors and information hazards (e.g. CBRN).

From the Code's Glossary, a **system prompt** is "a set of instructions, guidelines, and contextual information provided to a model before a user interaction begins".

The Purpose of the Requirement

To interpret Measure 7.1, point (4), I turn to the purpose of the requirement. I argue that its likely purpose is to help the AIO assess whether a Signatory's measures to mitigate systemic risk are appropriate. In other words, a spec helps the AIO understand whether a Signatory's mitigations match the systemic risks that the model poses.

I provide two arguments for this purpose. The first is that specs could provide information about a Signatory's mitigations and intentions to mitigate systemic risk. In particular:

- Specs themselves can be used as safety mitigations. Some providers use descriptions of principles to shape model behavior during training. Others use system prompts to shape behavior during deployment.
- Specs establish a baseline against which providers identify problematic deviations and decide to intervene with additional safety mitigations. For example, OpenAI rolled back an update to GPT-4o after observing sycophancy that conflicted with intended model behavior.

The second argument is that a spec is part of the Model Report, the purpose of which is to "[report] to the AI Office information about [the Signatory's] model and their systemic risk assessment and mitigation processes and measures" (Commitment 7). This information helps the AIO judge whether such processes and measures are appropriate (Commitment 5; Recital (a)).

Inclusion of Significant Intended Behavior

The overarching requirement in Measure 7.1, point (4) is that Signatories must describe in the spec how they intend their models to operate.[3]

[3] As an aside, it seems likely that the Code requires Signatories to have intentions of some sort (i.e., they cannot simply say that they have no intentions). The phrasing "*how* [emphasis added] Signatories intend the model to operate" presupposes that intentions exist. Moreover, the Code uses "if available" elsewhere to grant leeway when Signatories lack a particular item, but notably omits this qualifier in Measure 7.1, point (4). I do not expect

A potential problem is that Signatories might omit certain intended behaviors, whether by simple mistake or a lack of diligence in documentation. Some omissions could have a significant impact on the systemic risks posed by the model, and therefore impact the AIO's assessment of whether systemic risk mitigations are appropriate. For example, a former version of Claude's Constitution did not include any principles on model autonomy, which could affect loss of control risk.

What might be grounds on which to suspect such omissions? Indirect evidence of Signatories' intentions exists in other sources (e.g. the Safety and Security Framework or Model Report). For example:

- Model evaluations for sycophancy suggest that Signatories intend their models not to be sycophantic.
- Descriptions of how training data is generated and filtered could suggest traits that the Signatory wants to instill into the model.
- Ads about a model's ability to perform tasks autonomously suggest that the Signatory has intentions about model autonomy.

If the AIO suspects omissions, it should ask Signatories to confirm whether they were deliberate, pointing to alternative evidence of intentions as necessary.

How Thoroughly Must Principles be Specified?

Next, I interpret Measure 7.1, point (4)(a): specs must specify principles.

"Specify" means "to name or state explicitly in detail" (Merriam-Webster). The Code does not state a required level of detail. In practice, providers could specify principles at different levels of detail:

- A former version of Claude's Constitution includes the high-level principle "Please choose the response that most supports and encourages freedom, equality, and a sense of brotherhood", without further elaboration. The most recent version describes high-level principles like "being broadly ethical" in much more detail.
- The OpenAI Model Spec provides detailed descriptions, example behavior, and exceptions for all of its principles, such as the principle not to generate restricted content.
- As a middle ground, the Grok 4 chat system prompt[4] provides specific principles, such as disallowing queries intending to engage in "Conducting cyber attacks, including

this question to be controversial, especially since some major providers already describe intended model behavior (e.g. OpenAI Model Spec, Claude's Constitution).

[4] As the terms are typically used, a system prompt is not necessarily a model spec. The former is provided as input to the model, whereas the latter is a description of intended behavior that may or may not be provided as input. At the same time, a Signatory could potentially try to submit a system prompt to fulfill the spec requirement in the Code.

ransomware and DDoS attacks", but does not provide example behavior or detailed descriptions.

The Code likely requires principles to include all details that could have a significant impact on the systemic risk posed by the model. This is due to the purpose of Measure 7.1, point (4) discussed above: detailed specification of certain principles could be necessary for the AIO to assess whether the Signatory's measures to mitigate systemic risks are appropriate.

This requirement does not mean that high-stakes principles must always be described verbosely. Sometimes, a few words will be enough to convey the meaning of the principle. For example, the principle "Do not provide any information related to the construction of CBRN weapons, even general factual information" seems detailed enough. On the other hand, the principle "do not comply with requests for obtaining CBRN weapons" is ambiguous about whether the model is allowed to respond with general factual information. Such information could potentially facilitate malicious use, but might also have an educative value and aid defenders. Transparency about how Signatories handle this trade-off could help the AIO assess whether systemic risk mitigations are appropriate.

Principles that potentially deserve a greater level of detail include those relating to the specified systemic risks in Appendix 1.4 (e.g. loss of control risk).

Managing Significant Conflicts

Specs must state "how the model is intended to prioritize different kinds of principles and instructions" (Measure 7.1, point (4)(b)). Prioritisation is important because principles and instructions can conflict. For example, the principle "Try to prevent imminent real-world harm" could conflict with the principle "Be honest and transparent".

Yet, the Code does not specify which conflicts specs should focus on. Stating how the model is intended to handle all types of conflicts could be an extreme documentation burden. For example, the OpenAI Model Spec contains more than 30 principles, which amounts to more than 800 potential conflicts between pairs of principles, excluding principle-instruction and instruction-instruction conflicts.

I argue that the Code likely only requires specs to state how the model is intended to prioritize principles and instructions in case of significant potential conflicts between them, for the following reasons:

- The purpose of a spec is to help the AIO assess whether a Signatory's measures to mitigate systemic risk are appropriate.
- The most relevant conflicts for this purpose are those that could significantly affect the systemic risks posed by the model ("significant conflicts").

- For example, a prompt injection could subvert the user's instructions by instructing the model to manipulate the user.

One approach to fulfilling this requirement is to have an explicit chain of command among all principles and instructions, as in the OpenAI Model Spec.

Otherwise, the AIO could try to identify significant potential conflicts that are missing from a spec. Some potential conflicts to check include:

- "Be honest" vs. "avoid information hazards"
- User instructions vs. principles for preventing risky activities
- User instructions vs. external data sources (e.g. prompt injections)

System Prompts for All AI Systems

Lastly, Measure 7.1, point (4)(d) requires specs to include the system prompt, but is ambiguous about which system prompt must be provided. Indeed, the same model can be used with different system prompts. For example, Grok 4's chat system prompt differs from its API system prompt.

I argue that the Code requires specs to include system prompts for all of the Signatory's AI systems that integrate the model in question.[5]

The first reason is that providing only a single system prompt could hinder the AIO's ability to assess whether a Signatory's measures to mitigate systemic risk are appropriate. For example, Grok 4's chat system prompt includes a principle absent from its API prompt: to "pursue a truth-seeking, non-partisan viewpoint" when queries contain a "subjective political question forcing a certain format or partisan response". This additional principle could help to mitigate manipulation risks.

The second reason is that specs are a governance mitigation, by virtue of their inclusion in Measure 7.1. As a mitigation, they must take into account the model's integration into the Signatory's AI systems, as reasonably foreseeable. From Recital (b):

> "Signatories also recognise that the assessment and mitigation of systemic risks should include, as reasonably foreseeable, the system architecture, other software into which the model may be integrated [...] because of their importance to the model's effects, for example by affecting the effectiveness of safety and security mitigations."

[5] It seems unlikely that the Code requires sharing downstream developers' system prompts in general. Before placing a model on the market, providers cannot reasonably foresee the system prompts that downstream developers (or users for that matter) will provide to their AI systems. Furthermore, a downstream developer may not necessarily be classified as a provider of a GPAISR. If not, then the downstream developer need not comply with Article 55 of the AI Act, and therefore has little reason to sign on to the Code. See Section 3.2 of the Commission Guidelines for further details.

Separately, even if these arguments fail, the AIO could potentially request these system prompts under Measure 10.1 because they are likely a part of "a detailed description of how the model is integrated into AI systems" (Measure 10.1).

Some, but not all, providers already provide system prompts for their AI systems:

- xAI provides system prompts for the Grok 4 API, chat assistant, and X assistant.
- Anthropic provides system prompts for its web, Android, iOS apps, but not for the Anthropic API.[6]
- OpenAI and Google DeepMind have not provided system prompts.

Because the system prompt by definition is "provided to a model before user interaction begins", system prompts submitted to the AIO must match those in use. Inconsistencies could arise from routine updates or unauthorized changes, such as during the Grok 4 antisemitism incident.

To check whether submitted system prompts match those in use, the AIO could consider:

- Asking Signatories whether submitted system prompts are up to date.
- Checking whether third parties have extracted and released system prompts in use, while remaining cognizant of the potential unreliability of such evidence.
- Requesting additional documentation (Measure 10.1) of processes that prevent unauthorized changes to the system prompt, if they exist. Such processes could potentially count as safety mitigations (Measure 10.1, point (4)) because unauthorized changes could contribute to systemic risk (e.g. removal of an instruction to refuse CBRN topics).

Conclusion

Measure 7.1, point (4) of the Code requires Signatories to provide model specs in their Model Reports, but does not explicitly state:

- To what level of detail Signatories must specify principles;
- What it means to prioritize different kinds of principles and instructions; or
- Which system prompt(s) must be provided.

Based on the intended purpose of a spec, I argued the Code requires that specs

1. Include all intended behavior that could have a significant impact on the systemic risks posed by the model.

[6] Presumably, the reason is that the API does not give the model any system prompt. However, this point is not explicitly stated.

2. In specifying principles, include all details that could have a significant impact on the systemic risks posed by the model.
3. State how the model is intended to prioritize principles and instructions in case of significant potential conflicts between them.
4. Include system prompts for all of the Signatory's AI systems that integrate the model.

About the Authors

Alan Chan
Research Fellow, GovAI

Alan's research focuses on governing AI agents. He is also interested in technical AI governance and societal resilience more broadly. He obtained his PhD from Mila (Quebec AI Institute).

Acknowledgements

I am grateful to Jonas Freund, Sophie Williams, Aidan Homewood, Noemi Dreksler, Elias Groll, and Markus Anderljung for useful feedback on this piece. All errors remain my own.

About GovAI

GovAI is a 501c(3) non-profit organisation. Our mission is to help decision-makers navigate the transition to a world with advanced AI, by producing rigorous research and fostering talent. Researchers at GovAI work on a wide range of topics, with a particular emphasis on the security implications of frontier AI.