draft

General Purpose AI Code of practice

General Purpose AI Code of practice

Il Codice di condotta sull'IA di uso generale (General Purpose AI Code of Practice) è un documento che fornisce una guida ai fornitori di modelli di intelligenza artificiale di uso generale (GPAI) e di modelli di intelligenza artificiale di uso generale con rischio sistemico sui requisiti del Regolamento UE sull'intelligenza artificiale (AI Act). **Il codice mira a chiarire come i fornitori possono dimostrare la conformità, consentire all'AI Office di valutare la conformità e supportare la crescita di un ecosistema di IA sicuro e affidabile.**

Il codice è in fase di elaborazione e questa risposta si basa sulla seconda bozza, pubblicata a gennaio 2025. Il codice finale dovrebbe essere pronto entro il 2 maggio 2025.

Ecco alcuni punti chiave del Codice di condotta sull'IA di uso generale:

- **Trasparenza:** I fornitori di modelli GPAI devono fornire documentazione dettagliata sui loro modelli, comprese informazioni su formazione, test e potenziali rischi. Questo aiuta a garantire che gli utenti comprendano le capacità e i limiti dei modelli.
- **Copyright:** I fornitori di modelli GPAI devono attuare misure per rispettare la normativa sul copyright, in particolare per quanto riguarda l'utilizzo di dati protetti da copyright per l'addestramento dei modelli.
- **Rischio sistemico:** I fornitori di modelli GPAI con rischio sistemico devono adottare misure per valutare e mitigare i potenziali rischi sistemici posti dai loro modelli, come la diffusione di disinformazione o la creazione di armi autonome.
- **Governance:** Il Codice delinea i principi di governance per i fornitori di modelli GPAI, tra cui l'istituzione di quadri di responsabilità e la conduzione di valutazioni dei rischi.

Il Codice di condotta sull'IA di uso generale è uno strumento importante per promuovere lo sviluppo e l'utilizzo responsabile dell'IA. Fornendo una guida chiara ai fornitori, il Codice contribuisce a garantire che l'IA sia sviluppata e utilizzata in modo da rispettare i diritti fondamentali e i valori dell'UE.

Trasparenza

Ecco le principali linee guida sulla trasparenza delineate nel Codice:

Documentazione per l'AI Office e i provider downstream:
•I fornitori di modelli GPAI sono tenuti a redigere e mantenere aggiornata una documentazione tecnica completa che copra vari aspetti del loro modello. Questa documentazione deve essere messa a disposizione, su richiesta, dell'AI Office e delle autorità nazionali competenti.
•Una versione di questa documentazione, potenzialmente semplificata, deve essere fornita anche ai provider downstream, ovvero a coloro che intendono integrare il modello GPAI nei propri sistemi di IA.

Informazioni specifiche da includere nella documentazione:
La documentazione deve includere informazioni dettagliate su una vasta gamma di argomenti, tra cui:
•Informazioni generali sul modello, come nome, versione, famiglia e modalità di input e output.
•Dettagli sul processo di addestramento, test e validazione del modello, comprese le metodologie utilizzate, i dati utilizzati e le prestazioni ottenute.
•Informazioni sull'architettura del modello, le dipendenze hardware e software, le licenze e i termini di utilizzo.
•Misure di sicurezza e indicazioni sull'uso responsabile del modello.
•Informazioni relative al copyright, come l'identificazione e il rispetto delle riserve di diritti sui dati utilizzati per l'addestramento.

Trasparenza pubblica (incoraggiata, ma non obbligatoria):
•Il Codice incoraggia i fornitori a valutare se alcune informazioni nella documentazione possono essere divulgate pubblicamente per promuovere una maggiore trasparenza.
•Questo potrebbe includere informazioni sulle misure adottate per conformarsi alla normativa sul copyright, sui rischi potenziali del modello e sulle misure di mitigazione adottate.

Obiettivo della trasparenza:
L'obiettivo principale di queste linee guida è quello di fornire agli utenti e alle autorità le informazioni necessarie per comprendere appieno le capacità, i limiti e i potenziali rischi dei modelli GPAI. La trasparenza mira a promuovere un uso responsabile dell'IA e a facilitare l'identificazione e la mitigazione dei rischi, contribuendo a un ecosistema di IA più affidabile e sicuro.

Copyright

Ecco le principali linee guida per il copyright delineate nel Codice:

•**Politica interna sul copyright:** I firmatari del Codice si impegnano a **stipulare e attuare una politica interna per garantire il rispetto del diritto d'autore dell'UE** durante l'intero ciclo di vita dello sviluppo del modello GPAI. Questa politica deve coprire aspetti come l'acquisizione dei dati, l'addestramento, i test e l'immissione sul mercato.

- •Devono anche **rendere pubblica una sintesi di questa politica** per promuovere la trasparenza.
- •Inoltre, devono designare **un punto di contatto per la comunicazione con i titolari dei diritti** interessati, offrendo loro la possibilità di presentare reclami in caso di utilizzo non autorizzato delle loro opere.

•**Due diligence sul copyright:** Prima di utilizzare set di dati di terze parti per l'addestramento dei modelli, i firmatari devono eseguire una **due diligence sul copyright**. Ciò significa ottenere garanzie dai fornitori di dati sulla conformità al diritto d'autore e valutare la plausibilità di tali garanzie.

•**Rispetto del Robot Exclusion Protocol (robots.txt):** I firmatari devono utilizzare web crawler che rispettino le istruzioni contenute nei file robots.txt, un meccanismo che consente ai proprietari di siti web di indicare quali parti del loro sito possono essere scansionate.

•**Identificazione e conformità ad altre espressioni di riserva dei diritti:** I firmatari si impegnano a fare il possibile per identificare e rispettare altri metodi validi per esprimere riserve di diritti sui contenuti online, ad esempio, tramite metadati o altri standard tecnici.

•**Prevenzione dell'overfitting legato al copyright:** Per i modelli GPAI generativi (come quelli che generano testo o immagini), i firmatari devono adottare misure per prevenire l'overfitting, ovvero la situazione in cui il modello riproduce troppo fedelmente i dati di addestramento, rischiando di violare il diritto d'autore.

•**Divieto di usi del modello che violano il copyright:** I firmatari devono vietare esplicitamente gli usi del loro modello che violano il copyright, ad esempio, nella loro politica d'uso accettabile o nei termini di servizio.

•**Trasparenza sulla conformità al copyright:** I firmatari devono pubblicare informazioni chiare sulle misure che adottano per conformarsi al diritto d'autore, come l'elenco dei crawler utilizzati e le loro funzionalità robots.txt.

L'obiettivo finale di queste linee guida è garantire che i modelli GPAI siano sviluppati e utilizzati nel rispetto del diritto d'autore, promuovendo un ambiente di innovazione equo e sostenibile.
Si noti che queste linee guida si applicano a tutti i fornitori di modelli GPAI, con alcune eccezioni per le PMI e per i modelli rilasciati con licenza open source.

Rischi sistemici

Obiettivo principale: garantire che lo sviluppo e l'implementazione dei modelli GPAI con rischio sistemico avvengano in modo responsabile, riducendo al minimo i potenziali impatti negativi su salute pubblica, sicurezza, diritti fondamentali e la società nel suo insieme.
Ecco le linee guida chiave per i rischi sistemici delineate nel Codice:

1. Tassonomia dei rischi sistemici:
•Il Codice include una tassonomia dei rischi sistemici, che aiuta a identificare, analizzare e valutare i rischi potenziali posti dai modelli GPAI con rischio sistemico.
•Questa tassonomia non è esaustiva e viene costantemente aggiornata per riflettere i progressi scientifici, i cambiamenti sociali e l'emergere di nuovi rischi sistemici.
•I firmatari sono incoraggiati ad utilizzare questa tassonomia come punto di partenza per la propria valutazione del rischio e a considerare altri rischi pertinenti oltre a quelli elencati.

2. Framework di sicurezza e protezione:
•I firmatari si impegnano ad adottare, attuare e rendere disponibile all'AI Office un Framework di sicurezza e protezione (SSF).
•Questo framework dettaglia le politiche di gestione del rischio che il firmatario adotta per valutare e mitigare i rischi sistemici derivanti dai propri modelli GPAI.
•Il framework deve essere proporzionato ai rischi sistemici identificati e deve essere aggiornato regolarmente per riflettere le migliori pratiche e le conoscenze emergenti.

3. Valutazione e mitigazione del rischio lungo il ciclo di vita del modello:
•I firmatari si impegnano a valutare e mitigare i rischi sistemici durante l'intero ciclo di vita dello sviluppo e dell'implementazione dei loro modelli.
•Questa valutazione deve avvenire in punti chiave del ciclo di vita, tra cui prima dell'addestramento, durante l'addestramento, prima dell'implementazione, durante l'implementazione e dopo il ritiro del modello.

4. Raccolta di prove:
•I firmatari si impegnano a condurre una rigorosa raccolta di prove sui rischi specifici presentati dai loro modelli.
•Ciò include l'utilizzo di una serie di metodi, tra cui prove indipendenti dal modello e valutazioni del modello all'avanguardia.
•Le valutazioni del modello devono essere progettate per elicitare in modo appropriato le capacità e le propensioni del modello, riducendo al minimo il rischio di sottostima.

5. Mitigazioni tecniche:
•I firmatari si impegnano ad attuare mitigazioni tecniche proporzionate ai rischi identificati.
•Queste mitigazioni possono includere la modifica del comportamento di un modello, la limitazione dell'implementazione di un modello o la fornitura di contromisure o altri strumenti di sicurezza ad altri attori.
•I firmatari sono incoraggiati a considerare l'implementazione di mitigazioni di sicurezza per proteggere i pesi del modello non rilasciati e le informazioni algoritmiche associate.

Rischi sistemici

6. Decisioni di sviluppo e implementazione:
•I firmatari si impegnano a stabilire un processo per decidere se procedere o meno con lo sviluppo e l'implementazione di un modello GPAI con rischio sistemico.
•Questo processo deve tener conto dei risultati della valutazione del rischio e deve includere condizioni per non procedere se le mitigazioni sono insufficienti.
•I firmatari devono anche considerare di implementare un'implementazione graduale per mitigare i rischi durante la fase di lancio.

7. Mitigazioni della governance:
•I firmatari si impegnano ad allocare responsabilità e risorse adeguate in tutta l'organizzazione per valutare e mitigare i rischi sistemici.
•Ciò include l'istituzione di un chiaro processo decisionale e la promozione di una sana cultura del rischio.

8. Segnalazione di incidenti gravi:
•I firmatari si impegnano a istituire processi per identificare, documentare e segnalare incidenti gravi all'AI Office.
•Ciò include la definizione di cosa costituisce un "incidente grave" e l'allocazione di risorse sufficienti per indagare su qualsiasi sospetto di coinvolgimento del modello in un incidente grave.

9. Trasparenza pubblica:
•I firmatari sono incoraggiati, ma non obbligati, a fornire un livello appropriato di trasparenza pubblica sui loro sforzi di valutazione e mitigazione dei rischi.
•Questo potrebbe includere la pubblicazione dei loro SSF e dei Model Report, consentendo a ricercatori esterni e al pubblico di comprendere meglio i rischi sistemici e contribuire agli sforzi di mitigazione.
Le linee guida del Codice sui rischi sistemici sottolineano un approccio proattivo e completo alla gestione dei rischi associati ai modelli GPAI con capacità di alto impatto. L'obiettivo è promuovere un ecosistema di IA responsabile e affidabile, in cui i benefici dell'IA possano essere sfruttati riducendo al minimo i potenziali danni.

Rischi sistemici

Criteri di definizione per i rischi sistemici:
I rischi sistemici, come definiti nell'AI Act (articolo 3(65)), sono quei rischi specifici delle capacità di impatto elevato dei modelli di IA generici. Devono avere un impatto significativo sul mercato dell'Unione Europea a causa della loro portata o a causa di effetti negativi, reali o ragionevolmente prevedibili, sulla salute pubblica, sulla sicurezza, sui diritti fondamentali o sulla società nel suo complesso. Questi rischi possono propagarsi su larga scala attraverso la catena del valore.
Oltre alla definizione dell'AI Act, i documenti evidenziano ulteriori criteri per identificare i rischi sistemici, basati sulla loro natura e su considerazioni pratiche.

Criteri basati sulla natura del rischio:
•**Alta velocità:** il danno si materializza rapidamente, superando potenzialmente le misure di mitigazione esistenti.
•**Effetto composto o a cascata:** il danno si propaga a più livelli dei sistemi o della società.
•**Irreversibilità:** il danno è impossibile o molto difficile da invertire.
•**Impatto asimmetrico:** un piccolo gruppo di attori o un numero limitato di eventi causativi può avere un impatto significativo.

Criteri basati su considerazioni pratiche:
•**Copertura:** il rischio è riconosciuto in importanti framework e linee guida internazionali.
•**Valutabilità a livello dei fornitori di modelli:** esistono metodi di valutazione o possono essere ragionevolmente sviluppati.
•**Pratica consolidata:** il lavoro di alcuni fornitori di modelli di IA, della comunità scientifica o di altre entità ha dimostrato che gli aspetti del rischio possono essere valutati.

Rischi sistemici selezionati:
I documenti identificano specifici rischi considerati sistemici, tra cui:
•**Cyber offence:** rischi legati alle capacità offensive informatiche che potrebbero consentire attacchi informatici sofisticati su larga scala, anche su sistemi critici.
•**Manipolazione delle informazioni:** rischi derivanti dalla generazione di informazioni false o fuorvianti che potrebbero influenzare negativamente i processi democratici, la fiducia del pubblico o la sicurezza.
•**Perdita di controllo:** rischi associati alla perdita di controllo su modelli di IA potenti, in particolare se sviluppano capacità impreviste o agiscono in modi non allineati con l'intento umano.
•**Discriminazione su larga scala:** rischi di discriminazione illegale diffusa di individui, comunità o società, derivanti dall'utilizzo di modelli di IA in sistemi decisionali automatizzati ad alto rischio.

Rischi sistemici

Rischi aggiuntivi da considerare:
Oltre ai rischi selezionati, i documenti invitano i fornitori di modelli di IA a considerare anche:
•**Rischi relativi all'affidabilità dell'infrastruttura e del sistema**, come incidenti gravi, danni causati da malfunzionamenti o pregiudizi del modello, interferenze con infrastrutture critiche, danni derivanti dal controllo del modello su sistemi fisici e punti di guasto univoci a causa dell'eccessiva dipendenza da un piccolo numero di modelli.
•**Rischi relativi ai diritti fondamentali**, come violazioni della privacy e sorveglianza, nonché la generazione e diffusione di contenuti illegali o dannosi.
•**Qualsiasi altro rischio con effetti negativi su larga scala** sui diritti fondamentali, sulla salute e sicurezza pubblica, sui processi democratici, sulla sicurezza pubblica, sulla stabilità economica, sull'ambiente e sul benessere non umano o sull'agenzia umana.
Fonti di rischi sistemici:
I documenti descrivono tre categorie di fonti di rischio che possono contribuire ai rischi sistemici:
•**Capacità del modello:** capacità del modello che potrebbero essere utilizzate per causare danni, come capacità cyber-offensive, autonomia, capacità di auto-replicazione e auto-miglioramento.
•**Propensioni del modello:** caratteristiche del modello che potrebbero portare a comportamenti dannosi, come la disallineamento con l'intento umano, la tendenza a ingannare, il pregiudizio discriminatorio e la tendenza a "allucinazioni".
•**Affordances del modello e contesto di implementazione:** fattori esterni al modello che possono influenzare il rischio, come l'accesso a strumenti, le strategie di rilascio e il livello di supervisione umana.

La tassonomia dei rischi sistemici e i criteri per la loro definizione delineati nei documenti forniscono una guida per i fornitori di modelli di IA generici. Questa informazione li aiuta a comprendere, valutare e mitigare i potenziali rischi sistemici, contribuendo allo sviluppo e all'implementazione responsabile dell'IA.

Governance

Ecco le linee guida chiave per la governance delineate nel Codice:

1. Allocazione delle responsabilità per il rischio sistemico:
•**I firmatari si impegnano a garantire un'adeguata titolarità in materia di rischio sistemico a tutti i livelli organizzativi**, inclusi i livelli esecutivo e del consiglio di amministrazione.
•Ciò significa **individuare chiaramente gli individui o i team responsabili della valutazione, della mitigazione e della gestione dei rischi sistemici**.
•**Le responsabilità devono essere supportate da risorse adeguate** per garantire che i team abbiano il personale, il tempo e gli strumenti necessari per svolgere efficacemente i propri compiti.
•Per le **organizzazioni più grandi**, il Codice suggerisce di adottare modelli di governance del rischio consolidati, come il modello a tre linee di difesa. Questo modello prevede la separazione delle responsabilità per la gestione del rischio, il controllo del rischio e la supervisione del rischio, garantendo un sistema di controlli ed equilibri più robusto.

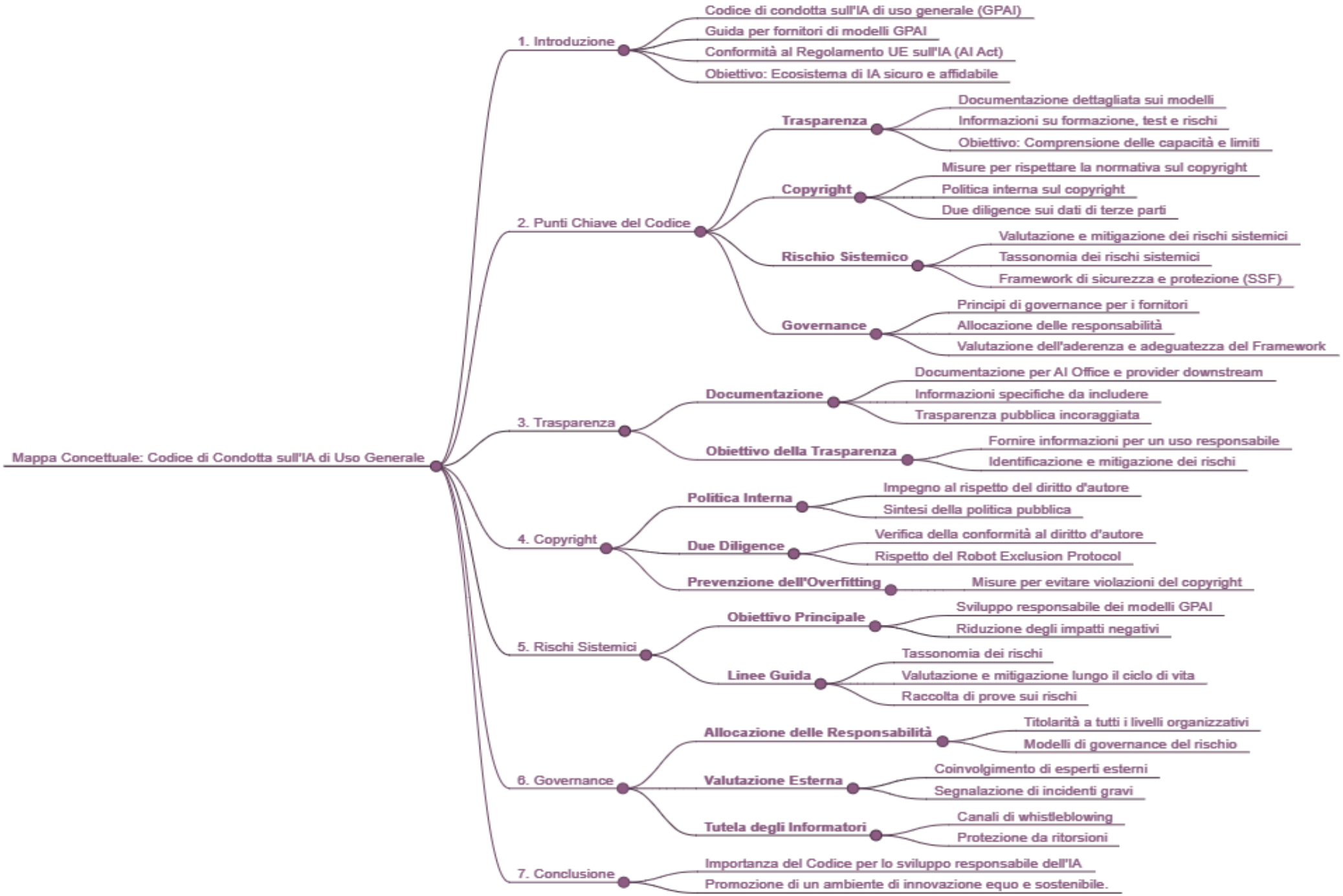
2. Valutazione dell'aderenza e dell'adeguatezza del Framework:
•I firmatari si impegnano a **valutare regolarmente l'aderenza e l'adeguatezza del proprio Framework di sicurezza e protezione (SSF)**.
•**L'aderenza** si riferisce alla misura in cui i processi e le procedure delineate nel Framework vengono effettivamente seguiti.
•**L'adeguatezza** si riferisce alla misura in cui il Framework è efficace nel mitigare i rischi sistemici identificati.
•Il Codice suggerisce di condurre queste valutazioni con cadenza regolare, ad esempio ogni sei mesi o dopo l'immissione sul mercato di un nuovo modello.
•I firmatari sono inoltre incoraggiati a coinvolgere **valutatori esterni qualificati** per condurre valutazioni indipendenti dell'aderenza e dell'adeguatezza del Framework.

3. Valutazione esterna del rischio:
•Il Codice incoraggia i firmatari a **coinvolgere esperti esterni per la valutazione dei rischi sistemici** in diverse fasi del ciclo di vita del modello.
•La **valutazione esterna può fornire una prospettiva indipendente e obiettiva**, contribuendo a identificare potenziali punti ciechi e a migliorare l'efficacia complessiva della gestione del rischio.
•I firmatari sono incoraggiati a collaborare con l'AI Office per **identificare valutatori esterni qualificati** e per stabilire protocolli chiari per le valutazioni esterne.

4. Segnalazione di incidenti gravi:
•I firmatari si impegnano a **istituire processi chiari per la segnalazione di incidenti gravi** all'AI Office.
•Questi processi devono includere una **definizione di cosa costituisce un "incidente grave"**, le procedure per la segnalazione degli incidenti e le misure da adottare in risposta a un incidente.
•I firmatari sono inoltre incoraggiati a **facilitare la segnalazione di incidenti da parte di terzi**, come utenti downstream o membri del pubblico.

Governance

5. Tutela degli informatori:
•I firmatari si impegnano a **implementare canali di whistleblowing e a fornire una protezione adeguata agli informatori** che segnalano potenziali violazioni del Codice o dell'AI Act.
•Ciò include la creazione di canali di segnalazione sicuri e confidenziali, la protezione degli informatori da ritorsioni e la garanzia di un'indagine imparziale sulle segnalazioni.
6. Trasparenza pubblica:
•Il Codice incoraggia i firmatari a **offrire un livello appropriato di trasparenza pubblica** sui loro sforzi di valutazione e mitigazione dei rischi.
•Ciò potrebbe includere la **pubblicazione di informazioni rilevanti sui loro SSF, Model Report e risultati delle valutazioni**, contribuendo a costruire la fiducia del pubblico e a promuovere la collaborazione con la comunità di ricerca.
In generale, le linee guida del Codice sulla governance mirano a promuovere un **approccio responsabile e proattivo** alla gestione dei rischi sistemici associati ai modelli GPAI con capacità di alto impatto. Si concentrano sulla creazione di un **solido sistema di responsabilità, supervisione e trasparenza** per garantire che i modelli GPAI siano sviluppati e implementati in modo sicuro ed etico.


Mappa Concettuale: Codice di Condotta sull'IA di Uso Generale
1. Introduzione
Codice di condotta sull'IA di uso generale (GPAI)
Guida per fornitori di modelli GPAI
Conformità al Regolamento UE sull'IA (AI Act)
Obiettivo: Ecosistema di IA sicuro e affidabile
2. Punti Chiave del Codice
Trasparenza
Documentazione dettagliata sui modelli
Informazioni su formazione, test e rischi
Obiettivo: Comprensione delle capacità e limiti
Copyright
Misure per rispettare la normativa sul copyright
Politica interna sul copyright
Due diligence sui dati di terze parti
Rischio Sistemico
Valutazione e mitigazione dei rischi sistemici
Tassonomia dei rischi sistemici
Framework di sicurezza e protezione (SSF)
Governance
Principi di governance per i fornitori
Allocazione delle responsabilità
Valutazione dell'aderenza e adeguatezza del Framework
3. Trasparenza
Documentazione
Documentazione per AI Office e provider downstream
Informazioni specifiche da includere
Trasparenza pubblica incoraggiata
Obiettivo della Trasparenza
Fornire informazioni per un uso responsabile
Identificazione e mitigazione dei rischi
4. Copyright
Politica Interna
Impegno al rispetto del diritto d'autore
Sintesi della politica pubblica
Due Diligence
Verifica della conformità al diritto d'autore
Rispetto del Robot Exclusion Protocol
Prevenzione dell'Overfitting
Misure per evitare violazioni del copyright
5. Rischi Sistemici
Obiettivo Principale
Sviluppo responsabile dei modelli GPAI
Riduzione degli impatti negativi
Linee Guida
Tassonomia dei rischi
Valutazione e mitigazione lungo il ciclo di vita
Raccolta di prove sui rischi
6. Governance
Allocazione delle Responsabilità
Titolarità a tutti i livelli organizzativi
Modelli di governance del rischio
Valutazione Esterna
Coinvolgimento di esperti esterni
Segnalazione di incidenti gravi
Tutela degli Informatori
Canali di whistleblowing
Protezione da ritorsioni
7. Conclusione
Importanza del Codice per lo sviluppo responsabile dell'IA
Promozione di un ambiente di innovazione equo e sostenibile.