**Responsible generative AI**

**Ansvarig användning av Generative AI**

Microsofts riktlinjer för **Responsible AI** fokuserar på att minimera risker och säkerställa etisk användning av **Generative AI**. Processen delas upp i fyra steg:

1. **Identify potential harms** – Identifiera möjliga risker och negativa konsekvenser.
2. **Measure harms** – Testa och kvantifiera dessa risker i systemets output.
3. **Mitigate harms** – Implementera åtgärder på flera nivåer för att minska skador.
4. **Operate responsibly** – Skapa en strukturerad **deployment & monitoring plan** för säker drift.

**Identifiera potentiella risker**

För att säkerställa en **Responsible AI**-lösning är det viktigt att identifiera **potential harms**. Exempel på risker är:

* **Offensive, discriminatory, or harmful content** – Output kan innehålla fördomsfulla eller stötande uttryck.
* **Factual inaccuracies** – AI kan generera felaktig information.
* **Encouraging unethical behavior** – Modellen kan oavsiktligt stödja olagliga eller farliga aktiviteter.

När risker identifieras behöver de **prioritiseras** baserat på sannolikhet och påverkan. Ett AI-system för matlagning har t.ex. större risk att ge felaktiga tillagningstider än att ge instruktioner för att skapa gift, men konsekvenserna av giftrecept är mycket allvarligare.

För att verifiera dessa risker kan **Red Teaming** användas, där experter aktivt försöker få AI:n att generera skadligt innehåll. Testresultaten dokumenteras och delas för att förbättra systemets säkerhet.

**Mätning av risker och skadlig output**

Efter att riskerna identifierats skapas en **baseline** genom att systemet testas med specifika **input prompts** för att utvärdera hur ofta skadligt innehåll produceras. Exempel på steg:

1. **Prepare prompts** – Använd noggrant designade frågor för att trigga skadlig output.
2. **Generate output** – AI:n genererar svar som analyseras.
3. **Categorize harm levels** – Utvärdera hur farlig outputen är (t.ex. "Safe", "Low", "Medium", "High").

Inledningsvis bör testning ske manuellt, men över tid kan **automated testing** användas för att skala upp valideringsprocessen.

**Minska risker genom flera skyddsnivåer**

För att minska skador används en **Layered Approach**, där åtgärder implementeras på fyra nivåer:

1. **Model layer** – Välj rätt AI-modell för uppgiften. För vissa applikationer kan en enklare modell minska risken för skadlig output.
2. **Safety System layer** – Inkludera **content filtering** och **abuse detection** i tjänster som **Azure AI Foundry** för att blockera olämpliga inputs och outputs.
3. **Metaprompt & Grounding layer** – Använd **prompt engineering** och **Retrieval Augmented Generation (RAG)** för att förbättra svarskvaliteten och minska desinformation.
4. **User Experience layer** – Användargränssnittet kan begränsa vilka frågor som ställs och ge tydlig information om AI:ns begränsningar.

**Ansvarig drift och monitorering**

Innan en **Generative AI**-lösning lanseras bör ett antal steg vidtas:

* **Pre-release reviews** – Säkerställ att lösningen följer **legal, security, privacy och compliance policies**.
* **Phased deployment** – Introducera AI:n stegvis för att identifiera eventuella problem innan full lansering.
* **Incident response & rollback plan** – Definiera hur skadliga incidenter hanteras och hur AI:n kan återställas vid behov.
* **User feedback & monitoring** – Möjliggör rapportering av **harmful, offensive eller inaccurate outputs** och samla in **telemetry data** för att övervaka användningen.

**Azure AI Content Safety** erbjuder inbyggd analys av skadligt innehåll, inklusive **prompt shielding, groundedness detection, och protected material detection**, för att ytterligare skydda mot missbruk.

Genom att följa denna process kan företag bygga **Responsible AI**-lösningar som minimerar risker och säkerställer etisk användning av **Generative AI**.