

Sûreté des IA – Anticiper et maîtriser les risques émergents

1 jour (7 heures)

Initiation

Présentiel ou classe à distance

à partir de

875 € HT par personne (TVA en vigueur) Repas inclus en présentiel

Contenu mis à jour le 25/09/2026. Document téléchargé le 07/10/2026.

Objectifs

A l'issue de cette formation, vous serez capable de :

- Identifier et catégoriser les risques émergents des Intelligences Artificielles (IA) à usage général, et définir la notion de "dilemme de la preuve"
- Evaluer l'efficacité et les limites des méthodes de gestion des risques actuelles
- Mettre en place un framework robuste pour des IA sûres
- Mettre en oeuvre une stratégie IA résiliente.

Les compétences visées par cette formation font l'objet d'une évaluation.

À qui s'adresse cette formation ?

Ingénieurs en sûreté de fonctionnement, chefs de projets, managers, ingénieurs en Intelligence Artificielle (IA), ingénieurs MLOps, Data Scientists, Data Analysts, développeurs IA, architectes Data et experts métier.

Prérequis

Avoir une familiarité avec les concepts de l'Intelligence Artificielle (IA) et de la sûreté de fonctionnement des systèmes.

Programme

1 Pendant votre formation

Sûreté des IA – Anticiper et maîtriser les risques émergents

Jour 1 – Matin

Comprendre le paysage des risques des IA à usage général (GPAI)

Introduction à la sûreté des IA

- Définition et distinction avec la sécurité (cyber) et l'éthique
- Introduction au concept de "General-Purpose AI"

Le dilemme de la preuve

- Explication du défi central pour les décideurs : l'évolution rapide des capacités vs la lenteur de l'émergence des preuves scientifiques
- Comment naviguer entre le risque d'une régulation prématurée et celui d'une vulnérabilité sociétale

Taxonomie des risques émergents

- Risques de malveillance : utilisation délibérée pour nuire
 - Contenu généré et criminalité : escroqueries, deepfakes, matériel pédopornographique généré par IA
 - Influence et manipulation : changement d'opinion, campagnes de propagande, dépendance psychologique
 - Cyberattaques : découverte de vulnérabilités, génération de code malveillant, automatisation des attaques
 - Risques biologiques et chimiques : assistance à la création d'armes, accès à des connaissances tacites
- Risques de défaillance : dysfonctionnements inattendus et non-intentionnels
 - Défis de fiabilité : hallucinations, échecs de raisonnement, vulnérabilité des agents IA
 - Perte de contrôle : scénarios où les systèmes agissent en dehors de tout contrôle (déception, évasion des garde-fous, "scheming")
- Risques systémiques : impacts à grande échelle sur la société et l'économie
 - Impacts sur le marché du travail : automatisation des tâches cognitives, effets sur l'emploi et les salaires, inégalités

Evaluer les méthodes de gestion des risques et leurs limites

Défis techniques et institutionnels

- Gaps scientifiques : manque de prévisibilité, boîte noire, écart entre l'évaluation pré-déploiement et la performance réelle ("evaluation gap")
- Asymétries d'information : manque de transparence des développeurs sur les données d'entraînement et les processus internes
- Défis institutionnels : rythme du développement vs rythme de la gouvernance, concurrence entre entreprises, fragmentation géopolitique

Jour 1 – Après-midi

Pratiques de gestion des risques actuelles

- Identification des risques : taxonomies, modélisation des menaces, prévisions ("forecasting")
- Analyse et évaluation : benchmarks, red-teaming, évaluations des capacités dangereuses
- Atténuation des risques : défense en profondeur, stratégies de déploiement (accès par API, déploiement progressif)
- Gouvernance des risques : Frontier AI Safety Frameworks, engagements "if-then", reporting d'incidents, transparence

Limites des approches actuelles

- Manque de standardisation
- Efficacité incertaine des garde-fous
- Difficulté d'évaluation des modèles open-weight
- Incapacité à anticiper les "inconnus inconnus"

Formuler et mettre en oeuvre une stratégie IA résiliente

Définir une stratégie de sûreté

- Comment passer de l'analyse à l'action, éléments clés :
 - Veille et évaluation : mettre en place une cellule de veille sur les capacités des modèles (benchmarks) et les incidents rapportés
 - Gouvernance interne : définir des seuils de tolérance au risque, allouer les responsabilités, créer un comité de sûreté
 - Contrôles techniques : appliquer une défense en profondeur (curation des données, filtres de contenu, surveillance des chaînes de pensée)
 - Gestion des dépendances : cartographier les risques liés à l'utilisation de modèles propriétaires (API) ou open-weight
 - Plan de résilience : élaborer un plan de réponse aux incidents

Les défis à venir

- Gérer l'incertitude
- S'adapter aux nouvelles capacités (agents IA)
- Maintenir un équilibre entre innovation et sûreté

Financement

Plusieurs dispositifs de financement sont accessibles via les OPCO (Opérateurs de Compétences), organismes agréés par le ministère du Travail dont le rôle est d'accompagner, collecter et gérer les contributions des entreprises au titre du financement de la formation professionnelle.

Pour plus d'information, une équipe de gestionnaires SKOLAE Formation spécialisée vous accompagne dans le choix de vos formations et la gestion administrative.

[Nous contacter pour en savoir plus →](#)

Modalités pédagogiques

Formation délivrée en présentiel ou distanciel* (blended-learning, e-learning, classe virtuelle, présentiel à distance).

Le formateur alterne entre méthode** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation).

Variables suivant les formations, les moyens pédagogiques mis en oeuvre sont :

- Ordinateurs Mac ou PC (sauf pour certains cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel)
- Environnements de formation installés sur les postes de travail ou en ligne
- Supports de cours et exercices

En cas de formation intra sur site externe à M2i, le client s'assure et s'engage également à avoir toutes les ressources matérielles pédagogiques nécessaires (équipements informatiques...) au bon déroulement de l'action de formation visée conformément aux prérequis indiqués dans le programme de formation communiqué.

* nous consulter pour la faisabilité en distanciel

** ratio variable selon le cours suivi

[Accessibilité aux personnes en situation de handicap](#)

Détails de la session

Le 17 décembre 2026 (1 jour / 7 heures)

Toulouse

Le 17 décembre 2026

Toulouse

- 17 décembre 2026 9h00 – 17h00