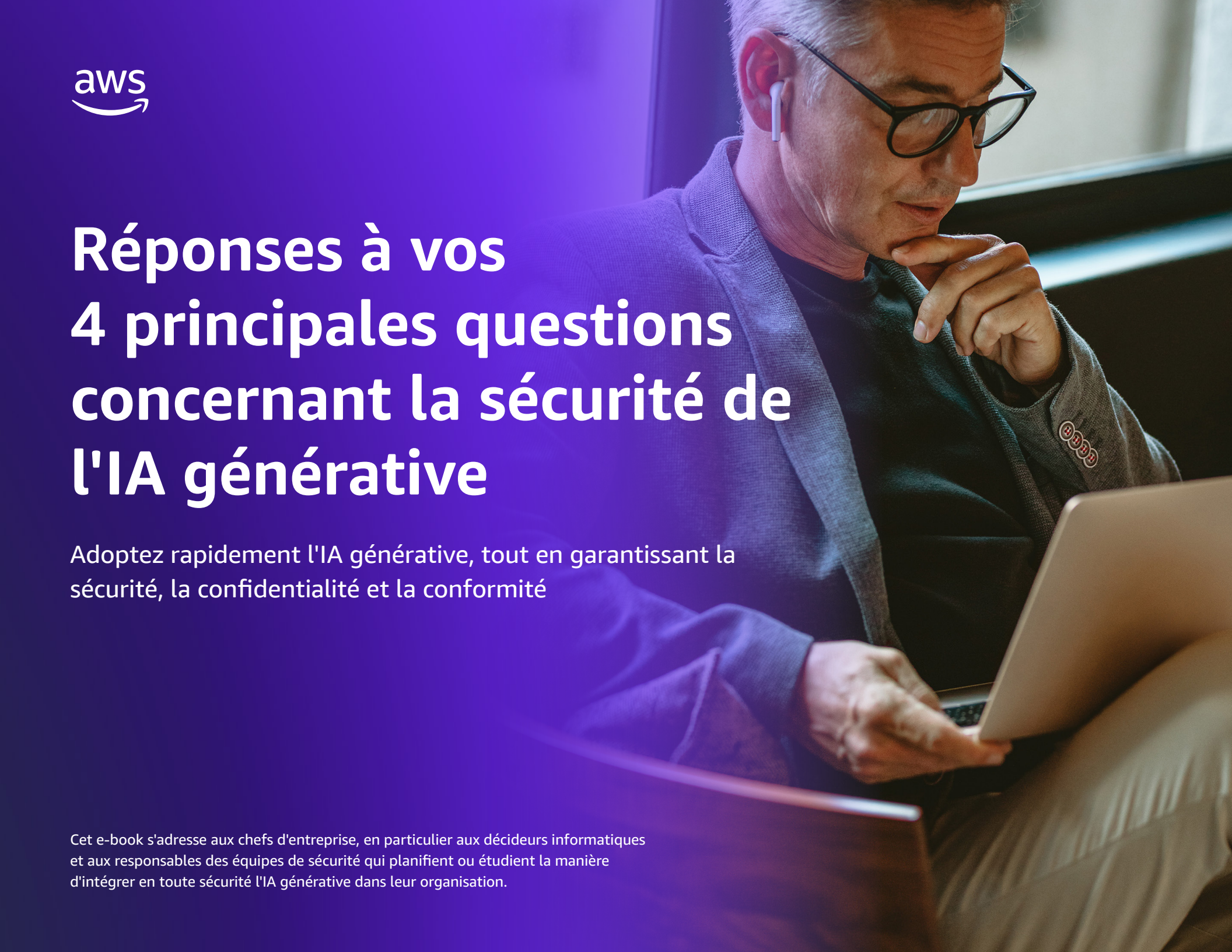


Réponses à vos 4 principales questions concernant la sécurité de l'IA générative

Adoptez rapidement l'IA générative, tout en garantissant la sécurité, la confidentialité et la conformité

Cet e-book s'adresse aux chefs d'entreprise, en particulier aux décideurs informatiques et aux responsables des équipes de sécurité qui planifient ou étudient la manière d'intégrer en toute sécurité l'IA générative dans leur organisation.

Table des matières

INTRODUCTION

À vos marques, prêts, générez : adoptez l'IA générative rapidement et en toute sécurité

La course à l'IA générative est lancée. Les entreprises s'emploient à réinventer les applications et l'expérience client, en raison des améliorations potentiellement très importantes qu'elles peuvent apporter à la productivité et à l'expérience.

Alors que l'ère de l'intelligence artificielle (IA) générative ne fait que commencer, les organisations en tirent déjà des avantages tangibles dans pratiquement toutes les unités opérationnelles. Cependant, les professionnels de la sécurité appellent à la prudence. Pour ces derniers, la confidentialité des données, les biais des modèles, la création de contenus nuisibles (tels que les hypertrucages) et les risques d'intervention malveillante sur les modèles sont autant de raisons d'aborder l'adoption de l'IA générative avec prudence.

Il est impératif que les organisations abordent l'IA générative avec une stratégie claire sur la façon de protéger leurs données, leurs utilisateurs et leur réputation, tout en permettant une adoption rapide et en améliorant l'expérience client.

Bien que cela constitue un défi à plusieurs facettes, les organisations doivent garder à l'esprit que les bonnes pratiques standard en matière d'IA, de machine learning, de protection des données et de sécurité des charges de travail dans le cloud s'appliquent toujours. En fait, votre organisation est peut-être mieux préparée à sécuriser l'IA générative que vous ne le pensez.

La mise en place de protections appropriées pour les charges de travail d'IA générative contribuera à stimuler l'innovation au sein de votre organisation. Cela donnera à vos équipes la confiance nécessaire pour concrétiser de grandes idées et la liberté de se concentrer sur le développement de votre entreprise.

Dans cet e-book, vous découvrirez 4 questions clés à vous poser au début de votre parcours vers des charges de travail d'IA générative plus sécurisées.

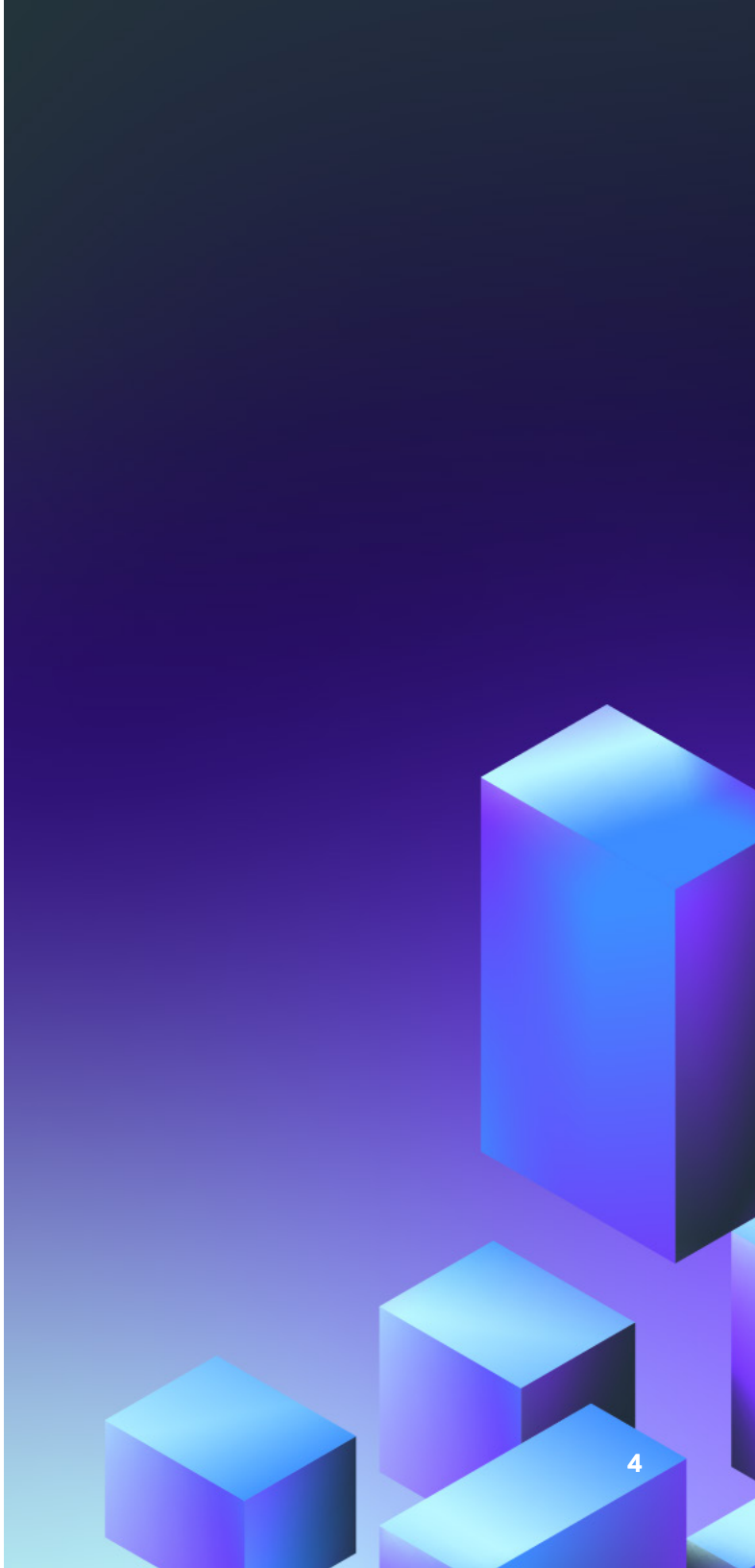
1. **Que devez-vous protéger ?**
2. **Comment pouvez-vous répondre aux problèmes de conformité ?**
3. **Comment pouvez-vous vous assurer que les modèles fonctionnent comme prévu ?**
4. **Par où devez-vous commencer ?**

Question 1 :

Que devez-vous protéger ?

Avant de pouvoir développer et déployer en toute sécurité des applications d'IA générative, il est important de comprendre ce qui doit être protégé. Il peut être utile de regrouper ces efforts en trois catégories :

- **Protection de vos charges de travail dans le cloud**
- **Protection de vos données**
- **Protection de vos applications d'IA générative**



Protection de vos charges de travail dans le cloud

Utiliser l'IA générative tout en respectant vos objectifs concernant la sécurité et la confidentialité commence par la protection de l'ensemble de votre infrastructure, de vos services et de vos configurations cloud. Pour ce faire, vous devrez d'abord faire la distinction entre vos responsabilités en matière de sécurité et celles qui sont assumées par votre fournisseur cloud.

Les clients Amazon Web Services (AWS) peuvent se référer au **modèle de responsabilité partagée** pour obtenir des conseils dans ce domaine. On y explique que, globalement, AWS est responsable de l'exploitation, de la gestion et du contrôle de l'infrastructure sur laquelle fonctionnent tous les services offerts dans le cloud AWS. C'est ce que l'on appelle la « sécurité *du* cloud ».

Les clients AWS, quant à eux, sont responsables de la gestion du système d'exploitation invité (notamment les mises à jour et les correctifs de sécurité), des logiciels d'application associés et de la configuration du pare-feu du groupe de sécurité fourni par AWS. L'étendue et les tâches spécifiques que les clients doivent accomplir dépendent des services AWS qu'ils choisissent d'utiliser. C'est ce que l'on appelle la « sécurité *dans* le cloud ».

Bien que la popularité de l'IA générative soit récente, les bonnes pratiques de sécurité traditionnelles restent un excellent point de départ. Cela inclut des pratiques d'hygiène de sécurité de base pour :

- Identity and Access Management (IAM)
- Détection et réaction
- Protection de l'infrastructure
- Protection des données
- Sécurité des applications

Protection de vos données

Vous devrez ensuite contribuer à garantir la sécurité et la confidentialité des données utilisées par vos applications d'IA générative. Cela peut inclure des informations exclusives, de la propriété intellectuelle (PI) de grande valeur et des informations personnelles identifiables (PII).

Les applications d'IA générative sont alimentées par des modèles de fondation (FM), qui sont entraînés sur de grandes quantités de données. Les FM analysent ces données pour identifier des modèles et apprendre à générer de nouveaux contenus similaires. Pour créer des applications d'IA générative qui répondent aux besoins spécifiques de votre entreprise, vous devrez généralement personnaliser un FM existant en l'entraînant à partir des données de votre organisation.

Pour protéger ces données, vous devrez prendre en compte les contrôles de confidentialité des données et les bonnes pratiques en matière de politique IAM.

Lorsque vous personnalisez un FM, assurez-vous que vos équipes travaillent avec une version du modèle qui soit stockée en toute sécurité et qui ne soit pas utilisée pour améliorer le FM lui-même. Si vous configurez une capacité dédiée à locataire unique dans **Amazon Bedrock**, le service peut associer ses instances d'inférence à votre **Amazon Virtual Private Cloud** (Amazon VPC) afin de lire et d'écrire sur **Amazon Simple Storage Service**(Amazon S3).

Un service IAM efficace permet de s'assurer que les bonnes personnes et les bonnes machines ont accès aux bonnes ressources dans les bonnes conditions. Le **cadre AWS Well-Architected** décrit les principes de conception et les bonnes pratiques architecturales qui contribuent à la gestion des identités. Cette ressource est un outil utile pour développer des politiques IAM et répondre à d'autres préoccupations en matière de sécurité, telles que la détection des menaces et la sécurité du réseau.

Protection de vos applications d'IA générative

Pour sécuriser l'IA générative au niveau des applications, vous devez en permanence identifier, classer, corriger et atténuer les risques. La première étape consiste à mettre en œuvre les bonnes pratiques existantes afin d'assurer la sécurité des environnements et des données.

À partir de là, vous devez réfléchir à la manière de placer la sécurité à un stade plus précoce du processus de développement. Cela peut contribuer à rationaliser vos efforts et permettre aux équipes de développement d'innover plus rapidement et avec une plus grande liberté, tout en évitant de faire de la sécurité un goulot d'étranglement.

Ensuite, vous devez réfléchir à la manière de protéger les trois éléments essentiels de toute application d'IA : les entrées, les sorties et le modèle lui-même.

Protection des entrées

Commencez par examiner les données qui entrent dans votre système d'IA. Les utilisateurs ne doivent pas avoir un accès direct au FM sans filtrage des entrées afin de réduire le risque d'attaques d'intégrité telles que la falsification, l'usurpation d'identité ou l'injection d'invites. Ces techniques d'attaque contournent les contrôles ou abusent du modèle. Les autres stratégies à envisager pour protéger les entrées sont l'automatisation de la qualité des données, la surveillance continue et la modélisation des menaces.

Protection des sorties

Les risques liés aux sorties des applications d'IA générative comprennent la divulgation d'informations, les incidents liés à la propriété intellectuelle et la mauvaise utilisation ou l'utilisation abusive du modèle pouvant nuire à la réputation de votre organisation. Lors de l'élaboration de votre modèle de menaces, tenez compte de l'empreinte des informations et du contexte d'utilisation, et incluez la détection et la surveillance de comportements complexes.

Protection du modèle lui-même

Enfin, il convient de prendre en compte la manière dont les adversaires peuvent tenter de supprimer des données du modèle lui-même ou de ses éléments associés. Les risques incluent les fausses représentations du monde réel ou des données dans le modèle, ainsi que les dommages causés à l'intégrité ou à la disponibilité du modèle. Modélisez les menaces qui pèsent sur les objectifs de votre entreprise et mettez en place une surveillance pour ces scénarios de menaces.

Question 2 :

Comment pouvez-vous répondre aux problèmes de conformité ?

En atténuant les risques liés à la conception et au développement d'applications d'IA générative, votre organisation peut renforcer la confiance de ses partenaires et de ses clients, préserver la réputation de sa marque et continuer à répondre à ses exigences en matière de conformité.

La réglementation légale en matière d'applications d'IA générative n'en est qu'à ses débuts. Il n'y a pas encore de consensus sur les bonnes pratiques. Ainsi, évoluer dans le labyrinthe des normes et des contrôles contradictoires entre les différentes juridictions représente un défi complexe et permanent.

Faites appel à vos conseillers juridiques et à vos experts en matière de confidentialité pour évaluer les exigences et les implications liées à la création de votre application d'IA générative. Il peut s'agir de vérifier vos droits légaux d'utiliser des données et des modèles spécifiques, mais aussi de déterminer l'applicabilité des lois sur la confidentialité, la biométrie, la lutte contre les discriminations et d'autres réglementations spécifiques à chaque cas d'utilisation.

Tenez compte des différences de législation entre les états, les provinces et les pays, ainsi que des nouvelles réglementations en matière d'IA qui sont proposées à travers le monde. Examinez ces considérations lors des phases suivantes de déploiement et d'exploitation.

Collaborer avec des pairs, des experts en IA et des organisations gouvernementales peut également vous aider à maintenir la conformité tout en montrant aux clients que vous prenez au sérieux les normes légales et éthiques en matière d'IA. Récemment, Amazon s'est joint à la Maison Blanche et à six grandes entreprises d'IA pour **s'engager volontairement à développer l'IA de manière responsable et sûre**, démontrant ainsi la valeur de tels engagements tout en établissant les bases d'une collaboration future.

Risques inhérents à l'intelligence artificielle

Comme toutes les solutions utilisant l'apprentissage automatique, les applications d'IA générative présentent des risques qui dépassent ceux des logiciels traditionnels. Afin de créer et de déployer en toute sécurité des applications utilisant l'IA générative, vous devrez étudier et élaborer des stratégies à même d'atténuer ces risques, notamment :

- Sorties biaisées, inexactes, trompeuses, nuisibles ou offensantes
- Complexités et coûts à l'échelle
- Jeux de données devenant trop volumineux, périmés ou éloignés de leur contexte d'origine
- Préoccupations concernant l'augmentation de l'opacité et de la reproductibilité
- Normes et procédures de test insuffisamment développées

Dans la section suivante, nous aborderons des stratégies générales visant à réduire certains de ces risques, ainsi que les bonnes pratiques permettant de définir les impacts professionnels, organisationnels et sociétaux de vos applications d'IA générative.

VISIBILITÉ DU COMPORTEMENT DU MODÈLE

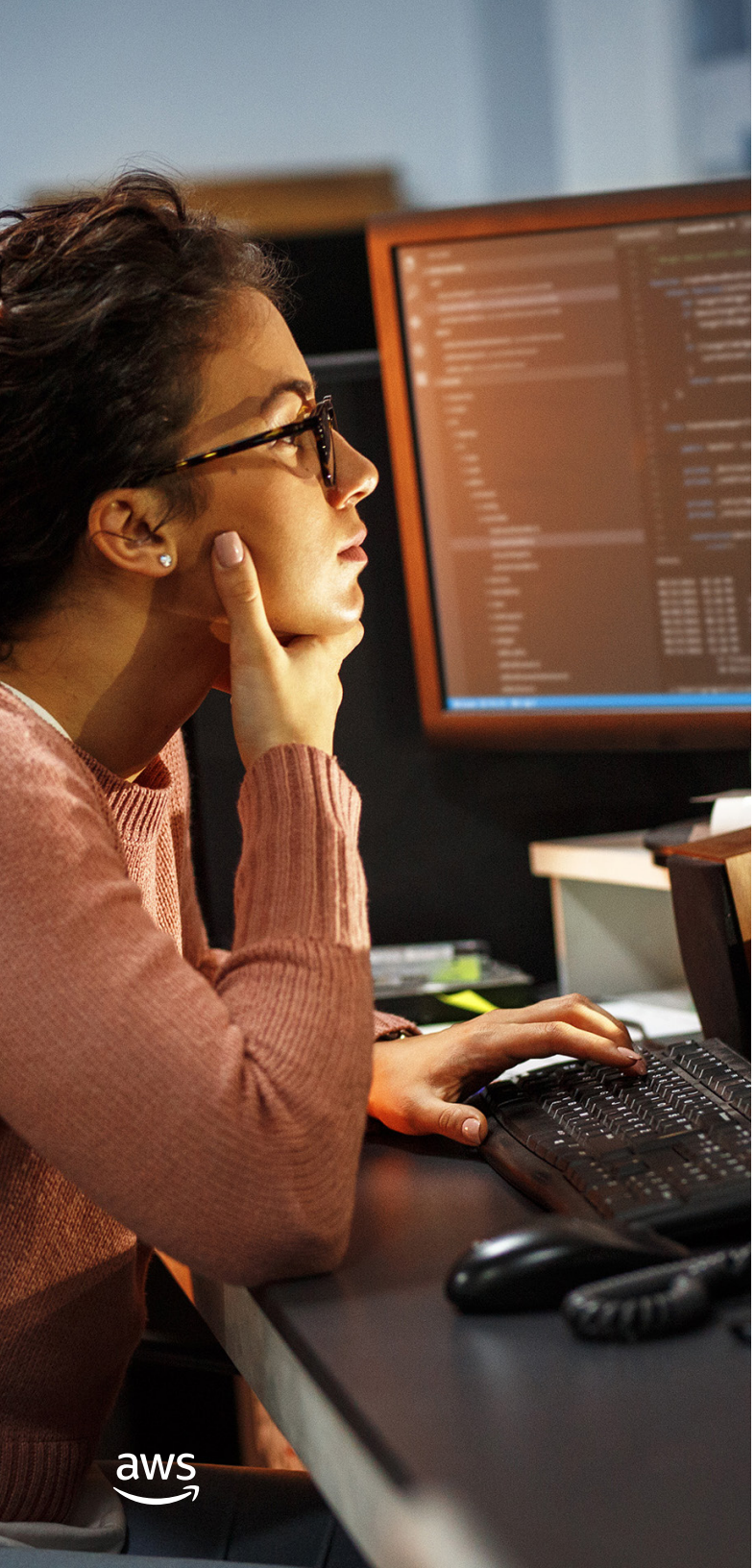
Question 3 :

Comment pouvez-vous vous assurer que les modèles fonctionnent comme prévu ?

Garantir une utilisation responsable de l'IA générative est devenu une tâche essentielle pour les entreprises, mais aussi un catalyseur essentiel de l'innovation continue.

Les FM s'entraînent sur des jeux de données volumineux et procèdent à des analyses complexes qui les aident à comprendre comment générer un contenu similaire. Bien que de nombreux FM produisent des résultats remarquables, le concept « garbage in, garbage out », ou GIGO, s'applique toujours. Si un FM reçoit des données inexactes, incomplètes ou biaisées en entrée, ses sorties peuvent présenter des défauts similaires.

Les données erronées ouvrent la voie à des abus, à des actions malveillantes et à d'autres risques. Au fur et à mesure que votre application d'IA générative se développe en termes d'utilisateurs, de portée et de fonctions, l'impact potentiel de ces problèmes augmente.



Promotion d'une IA responsable

S'engager dans une stratégie d'IA responsable vous aidera à faire face à ces risques. L'explicabilité, l'équité, la gouvernance, la confidentialité, la sécurité, la robustesse et la transparence font partie des dimensions d'une l'IA responsable. Comprendre la manière dont les différentes cultures et les différents groupes démographiques sont perçus, traités et impactés par l'application est également important.

Il est préférable de commencer à envisager l'IA responsable dès le début de votre parcours d'IA générative, puis de continuer à en tenir compte tout au long du cycle de vie de l'application en tant qu'élément clé de votre vision. Commencez par des actions relativement modestes et simples. Ensuite, évaluez l'impact de l'IA responsable sur votre conception, votre développement et vos opérations au fil du temps.

Lorsque vous rédigez des politiques responsables en matière d'IA et de gouvernance, réfléchissez à la manière dont votre application d'IA générative affectera vos utilisateurs, vos clients, vos employés et la société. Veillez à aborder l'équité des algorithmes, la représentation de la diversité et de l'inclusion, ainsi que la détection des biais.

Lutte contre la toxicité

La toxicité dans les grands modèles de langage (LLM) fait référence à la génération de textes grossiers, irrespectueux ou déraisonnables. Il existe de nombreuses stratégies pour prévenir la toxicité et garantir l'équité dans vos applications d'IA générative. Par exemple, vous pouvez identifier et supprimer les propos offensants ou les phrases biaisées de vos données d'entraînement. Vous pouvez également effectuer des tests d'équité plus précis, axés sur le cas d'utilisation spécifique de votre application, sur les publics cibles ou sur les invites et les requêtes les plus susceptibles d'être reçues.

Vous pouvez également entraîner des modèles de barrière de protection sur des jeux de données annotés qui identifient différents types et degrés de toxicité. Cela peut aider le FM à apprendre à détecter et à filtrer le contenu indésirable dans les données d'entraînement, les invites d'entrée et les sorties générées de manière automatisée.

Protection de la confidentialité

Plusieurs mesures permettent d'éviter l'exposition involontaire d'informations sensibles, de secrets commerciaux et de propriété intellectuelle lors de l'utilisation d'applications d'IA générative.

La suppression du modèle est une méthode qui contribue à répondre aux préoccupations en matière de confidentialité. Cela implique d'éliminer les données utilisées de manière inappropriée dès qu'elles sont identifiées, ce qui permet de supprimer les effets de ces données sur n'importe quel élément du FM.

Le partitionnement est une autre approche qui consiste à diviser les données d'apprentissage en portions plus petites et à utiliser ces dernières pour entraîner des sous-modèles distincts qui seront ensuite combinés pour composer le FM. Cette pratique peut simplifier considérablement les mesures correctives à l'égard des FM exposés à des informations privées ou risquant de l'être. Plutôt que de réentraîner l'ensemble du modèle, il vous suffit de supprimer les données indésirables ou utilisées de manière inappropriée de leur partition, puis de réentraîner ce sous-modèle.

Le filtrage et le blocage peuvent également être des approches efficaces. Ces méthodes comparent explicitement les informations protégées au contenu généré avant que l'utilisateur ne le voie. Si les deux sont trop similaires, le contenu est supprimé ou remplacé pour éviter toute exposition. Il peut également s'avérer utile de limiter le nombre de fois qu'un élément de contenu spécifique apparaît dans les données d'entraînement.

Amélioration de l'explicabilité et de l'auditabilité

Pour favoriser davantage l'IA responsable, il convient d'expliquer la méthodologie et les facteurs clés qui influencent les sorties de votre application. L'auditabilité est un autre élément important d'une IA responsable. Mettez en place des mécanismes qui vous permettent de suivre et d'examiner le développement et le fonctionnement de votre application d'IA générative. Cela vous aidera à identifier l'origine des problèmes et à répondre aux exigences en matière de gouvernance.

Pensez à documenter les décisions et les apports pertinents en matière de conception tout au long du cycle de développement. L'établissement d'un enregistrement traçable peut aider les équipes internes ou externes à évaluer le développement et le fonctionnement de votre application d'IA générative.

Rester responsable

Enfin, réfléchissez à la manière dont vous contribuerez à garantir l'adhésion continue à vos politiques d'IA responsable. Veillez à appliquer les leçons apprises et l'expérience acquise afin de faire évoluer vos pratiques en matière de sécurité et de confidentialité. Sensibilisez régulièrement tous les employés de votre organisation à leurs obligations concernant des pratiques d'IA générative sûres et sécurisées. Favorisez une culture de l'IA responsable, utilisez les bons outils pour vous aider à surveiller les performances des modèles et à informer sur les risques, et autorisez vos équipes à inspecter le modèle et ses éléments si nécessaire. Testez, testez et, en cas de doute, testez à nouveau.

Question 4 :

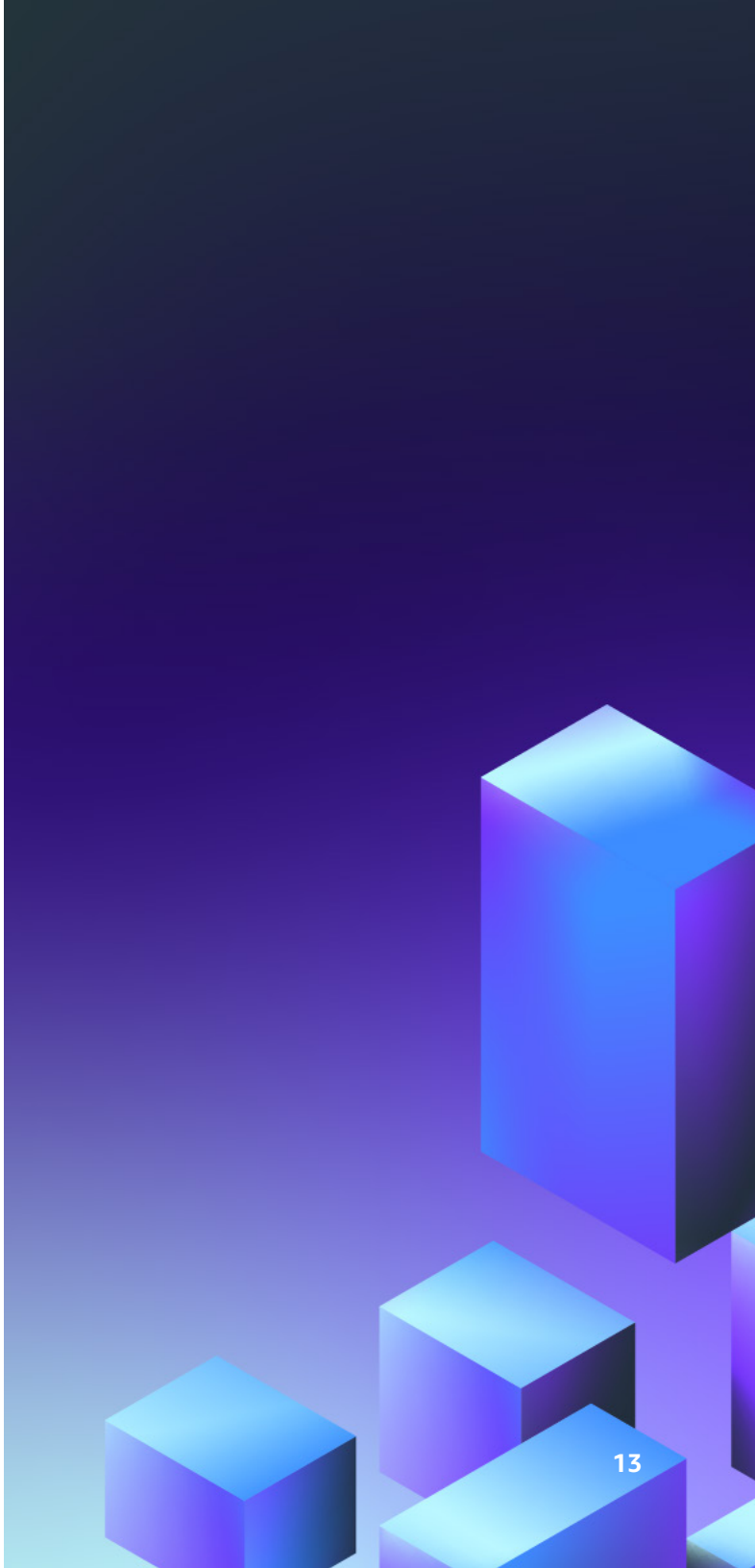
Par où devez-vous commencer ?

Sécuriser les applications d'IA générative n'est pas une mince affaire, et il n'existe pas d'ensemble universel de mesures à prendre pour y parvenir. Cependant, lorsque vous travaillez avec le bon fournisseur et que vous déployez les bons outils, la voie du succès devient beaucoup plus claire.

L'utilisation d'**Amazon Bedrock**, par exemple, peut considérablement simplifier et accélérer votre processus de développement d'applications d'IA générative sécurisées. Amazon Bedrock est un service entièrement géré qui met à disposition les FM d'Amazon et des principales start-ups d'IA via une API.

Lorsque vous personnalisez un modèle avec Amazon Bedrock, le service peut ajuster le modèle pour une tâche particulière sans que votre équipe ait à annoter d'importants volumes de données. Ensuite, Amazon Bedrock produit une copie séparée du FM de base auquel vous seul pouvez accéder et entraîne cette copie privée du modèle. Aucune de vos données n'est utilisée pour entraîner les modèles de base d'origine, ce qui permet de préserver la confidentialité et la sécurité de vos données propriétaires.

Vous pouvez également configurer vos paramètres **Amazon VPC** pour accéder aux API d'Amazon Bedrock et fournir à votre modèle des données de réglage fin de manière sécurisée. Vos données sont toujours chiffrées en transit et au repos au moyen de clés gérées par le service. De plus, avec **AWS PrivateLink**, vous pouvez transmettre vos données du cloud AWS à Amazon Bedrock exclusivement via le réseau AWS, jamais via Internet.



Amélioration de la confidentialité avec AWS

Que vous développiez des applications d'IA générative à l'aide d'Amazon Bedrock, d'un autre service (tel qu'**Amazon SageMaker**) ou de vos propres outils, lorsque vous exécutez et gérez vos applications sur AWS, vous bénéficiez de protections et de contrôles de la confidentialité à la pointe du secteur.

AWS prend en charge 143 normes de sécurité et certifications de conformité, ce qui contribue à satisfaire les exigences de nos clients dans le monde entier. Toutes vos données peuvent être chiffrées au repos à l'aide de vos propres clés **AWS Key Management Service** (Amazon KMS), ce qui vous permet de disposer d'un contrôle et d'une visibilité absolus sur la manière dont vos données et vos FM sont stockés et consultés.

CONCLUSION

Étapes suivantes

AWS s'engage à vous aider à créer des applications d'IA générative qui contribuent à la croissance de votre entreprise tout en vous aidant à atteindre vos objectifs de sécurité, de confidentialité et de conformité.

Nous sommes fermement convaincus que les applications d'IA générative peuvent être conçues, développées et exploitées en toute sécurité. Nous reconnaissons également la pertinence des préoccupations concernant la sécurité et la confidentialité liées à ces technologies. **L'IA générative crée de nouveaux défis** dans la définition, la mesure et l'atténuation des problèmes liés à la confidentialité des données, à la propriété intellectuelle, à la surveillance légale, à l'égalité et à la transparence.

Avec l'introduction de nouveaux produits, la complexité et l'échelle croissantes des solutions, les nouveaux paramètres d'entraînement et les jeux de données toujours plus volumineux, la sécurité de l'IA générative sera encore plus essentielle à l'avenir. En élaborant dès maintenant une stratégie de sécurité efficace et complète pour les charges de travail d'IA générative, vous pouvez maximiser votre avantage concurrentiel et vous préparer à l'avenir qui se profile de manière imminente.

Bonne nouvelle : les contrôles de base nécessaires pour concevoir, développer et exécuter en toute sécurité des applications d'IA générative sont en place depuis plusieurs années, et sont alignés sur des principes de sécurité du cloud fiables et éprouvés, à l'image de ceux que propose le **cadre AWS Well-Architected Framework**.

En explorant les pratiques décrites dans cet e-book, vous avez déjà fait un premier pas vers la sécurisation de vos charges de travail d'IA générative.

Passez maintenant à l'étape suivante avec AWS. Nous vous proposons des informations approfondies et des conseils spécifiques nécessaires pour rester au fait des sujets émergents, réfléchir à vos défis uniques et exploiter tous les avantages de l'IA générative, tout en protégeant vos données, vos clients et votre entreprise.

En savoir plus sur l'IA générative sur AWS ›

Démarrez rapidement avec Amazon Bedrock ›

Créez et personnalisez des FM sur Amazon SageMaker ›

Renforcez votre sécurité dans le cloud avec AWS ›

IA responsable : passez de la théorie à la pratique ›