GuiA

Guia para a Inteligência Artificial

GUIA PARA UMA INTELIGÊNCIA ARTIFICIAL
ÉTICA, TRANSPARENTE E RESPONSÁVEL
NA ADMINISTRAÇÃO PÚBLICA

VALORES, PRINCIPIOS E RECOMENDAÇÕES

ama AGÊNCIA PARA A MODERNIZAÇÃO ADMINISTRATIVA
tic.gov.pt
Cofinanciado por: COMPETE 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

Este documento resume conteúdos que integram o GuIA para a Inteligência Artificial Ética, Transparente e Responsável na Administração Pública, elaborado pela AMA e visa constituir uma leitura mais simples e focada nos

Valores, Princípios e Recomendações 1/3

PENSAR NUMA IA COM VALORES E PRINCÍPIOS É ACIMA DE TUDO PENSAR NUMA INTELIGÊNCIA ARTIFICIAL RESPONSÁVEL.

VALORES E PRINCÍPIOS

- Respeitados por todos os atores durante o ciclo de vida dos sistemas com IA;
- Promovidos por uma avaliação e evolução contínua das leis, dos regulamentos e das várias diretrizes;
- Alinhados com objetivos de sustentabilidade social, política, ambiental, educacional, científica e económica.

Os valores desempenham um papel importante como ideais que **motivam a orientação de medidas políticas e normas jurídicas**, dirigidas aos cidadãos e às empresas. A identificação de valores associados à IA, permite inspirar comportamentos desejáveis e representa os fundamentos dos diversos princípios.

> Sistemas com IA oferecem benefícios que podem ser compartilhados por toda a sociedade

Com o crescimento da dimensão das bases de dados, das novas tecnologias e dos mecanismos automáticos de recolha e partilha de dados, as primeiras iniciativas para a **proteção e privacidade de dados** ganharam relevo. No caso europeu, entrarem em vigor em maio de 2018.

Avanços na **legislação e na proteção de dados** precisam ser implementados, no alinhamento com os valores e princípios reconhecidos.

No caso de Portugal estes valores e princípios devem considerar:

A **Constituição Portuguesa**, nomeadamente:

- A dignidade da pessoa humana;
- Uma sociedade livre, justa e solidária.

A **Constituição Europeia**, nomeadamente:

- Os direitos individuais;
- As liberdades individuais.

E a **Declaração Universal dos Direitos Humanos**, nomeadamente:

- O direito à vida;
- O direito à segurança.

VALORES E PRINCÍPIOS BASILARES A UMA IA RESPONSÁVEL E TRANSPARENTE

- Inovação de forma responsável;
- Promoção de um ecossistema digital;
- Cooperação entre organismos para uma IA de confiança;
- Incorporação de feedback do sector privado, da indústria, de universidades e de organismos da AP;
- Promoção da partilha de dados em dados.gov.pt;
- Identificação de consequências negativas não intencionais que sistemas e soluções podem ter sobre os indivíduos e as comunidades a que se dirigem;
- Partilha de modelos de IA transparentes e explicáveis;
- Acesso aos consequentes benefícios;
- Identificação clara dos serviços que os fornecedores de IA disponibilizam;
- Promoção de um governo orientado para o utilizador, a abertura, a colaboração e a acessibilidade;
- Utilização inovadora e responsável das novas tecnologias;
- Disponibilização de ferramentas aos serviços públicos;
- Alinhamento dos sistemas de IA com elevado grau de autonomia com os valores humanos em toda a sua operação;
- Proteção prioritária dos valores sociais, de justiça e o interesse público;
- Benefício e capacitação do maior número de pessoas possível;
- A distribuição de prosperidade económica criada pela IA;
- Respeito e melhoria dos processos sociais e cívicos;
- Objeção às tecnologias e sistemas de IA utilizados para a tomada de decisões e/ou criação de serviços e produtos antagónicos aos valores prevalecentes na nossa sociedade, a exemplo, para o fabrico de armas, comércio de droga, prostituição, exploração infantil, tráfico humano, pornografia, e outros.

Os benefícios da IA devem colocar:

- Em 1.º lugar a Humanidade;
- Em 2.º lugar o Estado;
- E, por último, a **Organização**.

FIGURA 1: ELEMENTOS A TER EM CONTA NA GARANTIA DE UMA IA INCLUSIVA E QUE CONSIDERE TODOS NA SOCIEDADE.

O **bem-estar individual** e coletivo deve ser um fim da IA. Sistemas com IA podem ser utilizados para:

- Favorecer a habitação, cultura, educação e saúde;
- Promover a inclusão de populações sub-representadas;
- Reduzir as desigualdades económicas, sociais e de género, e ainda
- Produzir resultados benéficos para o planeta, contribuir para a ação climática e conservar os ambientes naturais.

O tratamento **igualitário e justo** de todos os indivíduos e grupos necessita que a conceção dos sistemas com IA e o tratamento de dados **considere conscientemente** a autonomia individual, grupos vulneráveis, princípios de não discriminação e um sentido de solidariedade.

Um requisito essencial para aumentar a **confiança** é que em todo o ciclo de vida, os sistemas com IA estejam sujeitos a **monitorização** do governo, empresas privadas, sociedade civil e outras partes interessadas independentes.

RECOMENDAÇÕES

No âmbito do cumprimento e preservação de valores e princípios que definirão uma IA responsável, identificaram-se e organizaram-se 10 categorias principais de recomendações, descritas infra em detalhe:

O Controlo Humano

- Controlo humano da tecnologia;
- Controlo humano dos algoritmos;
- Inclusão de rotinas para treino e validação dos mecanismos de aprendizagem;
- Revisão humana de decisões automatizadas - As pessoas devem ser governadas por pessoas;
- Capacidade de reverter decisões automatizadas.

A Transformação Digital e Tecnológica

- Investimentos em dados e dados abertos;
- Esforços para acelerar a digitalização da AP;
- Inclusão digital;
- Ações da academia, indústria e sociedade civil.

A Cooperação e Envolvimento

- Empresas de tecnologia de IA;
- Fornecedores de IA para a AP;
- Organizações de utilizadores finais do setor privado;
- Consultoria;
- Especialistas em ética aplicada à IA, tecnologias emergentes e em ciências sociais;

- Equipas multidisciplinares com uma variedade de valências;
- Apoio a start-ups.

A Justiça e Não discriminação

- Prevenção de preconceitos subjacentes aos dados;
- Prevenção de preconceitos subjacentes aos princípios e pressupostos;
- Inclusão no desenho das soluções;
- Inclusão no impacto das soluções;
- Dados representativos;
- Dados de elevada qualidade;
- Reduzir o impacto negativo para os funcionários e, quando viável, permitir a sua participação na conceção e implementação desses sistemas.

A Privacidade

- Controlo de dados do utilizador;
- Consentimento;
- Recomendação e informação de leis de proteção de dados;
- Capacidade de restringir o processamento;
- Direito à retificação;
- Direito de apagar registo.

A Promoção de Valores Humanos

- Foco na criação de benefícios para a sociedade;
- Refletir os Valores e a Ética do Setor Público, bem como as obrigações internacionais e direitos humanos;
- Acesso à tecnologia para todos;
- Acesso à informação para todos;
- Replicabilidade por indivíduos nas mesmas circunstâncias.

A Responsabilidade Profissional

- Colaboração entre atores e stakeholders;
- Design responsável;
- Consideração de efeitos a longo prazo;
- Integridade e excelência científica;
- Precisão por meio de análises aprofundadas em todas as etapas;
- Supervisão externa;
- Qualificar e certificar fornecedores.

A Responsabilização

- Recomendação para novos regulamentos;

- Avaliação de impactos mensuráveis;
- Requisitos para avaliação e auditoria;
- Verificabilidade e replicabilidade;
- Responsabilidade legal;
- Capacidade de intervir;
- Responsabilidade ambiental;
- Criação de órgão de monitorização;
- Correção de decisões automatizadas;
- Explicação satisfatória e auditável da ocorrência de resultados erróneos.

A Segurança e Proteção

- Mecanismos de confiabilidade;
- Mecanismos de previsibilidade;
- Geração de alarmística;
- Colaboração estreita do Governo com os técnicos e investigadores para investigar, prevenir e mitigar os potenciais usos maliciosos de IA.

A Transparência e Explicabilidade

- Análise de dados recorrendo a código aberto;
- Algoritmos de código aberto;
- Notificação dos utilizadores e/ou beneficiários ao interagirem com IA;
- Notificação quando um sistema de IA toma uma decisão sobre um indivíduo ou um grupo de indivíduos;
- Requisito de relatórios regulares ao longo de todo o ciclo de vida da solução de IA;
- Direito à informação por parte dos utilizadores e ou beneficiários;
- Aquisição transparente de tecnologia para o Governo;
- Acesso à explicação das decisões tomadas;
- Comunicação à comunidade das mudanças permitidas pela IA.

Numa perspetiva geral da conceção de soluções de IA, é recomendado às partes interessadas:

1. A criação de um Comité Ético e de um Comité de Especialistas, que inclua profissionais das áreas em que são utilizadas tecnologias de IA (por exemplo, um painel de médicos e um painel de juízes);
2. A escolha de uma metodologia de gestão de projetos adequada, alinhada com uma estratégia de comunicação com as partes interessadas e ajustada às expectativas previstas e à sua mudança;
3. A inclusão de programas de formação/qualificação dos recursos humanos e dos utilizadores/beneficiários;

4. O desenho de um *roadmap* em torno de IA Responsável considerando os seguintes aspetos:

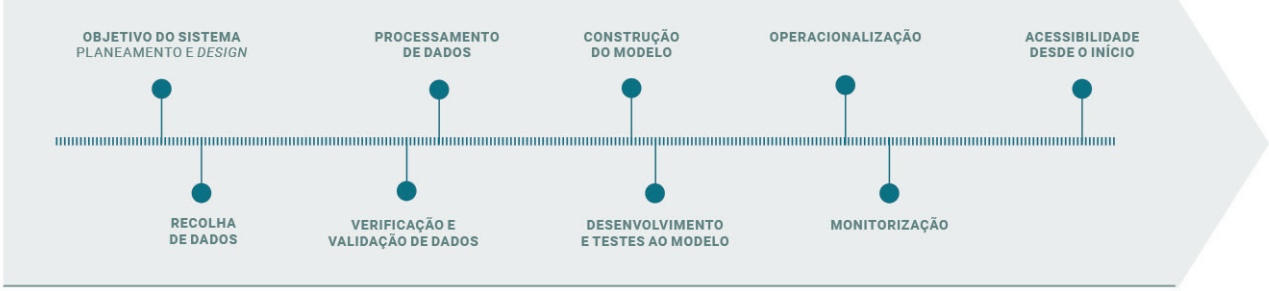

FIGURA 2: ETAPAS DE DESENVOLVIMENTO DE UM SISTEMA DE IA RESPONSÁVEL.

5. O planeamento de um projeto que responda às seguintes questões "Como?":
   - Inovar de forma responsável;
   - Identificar possíveis implicações éticas do uso da IA na organização e na sociedade;
   - Promover um ecossistema digital para IA;
   - Permitir a cooperação entre organismos para promover uma IA de confiança;
   - Incorporar pareceres do setor privado, indústria, academia, organismos da AP e promover a partilha de dados em dados.gov.pt;
   - Identificar consequências negativas não intencionais que os seus sistemas e soluções possam ter sobre os indivíduos e as comunidades que afetam;
   - Partilhar modelos de IA transparentes e explicáveis;
   - Gerar benefícios indiretos e identificá-los;
   - Tornar o Governo mais eficiente e melhor prestador de serviços;
   - Identificar fornecedores de serviços de IA com experiência comprovada e expertise em ética.

6. A consideração dos seguintes elementos, em relação aos algoritmos gerados:
   - Os pressupostos em que são baseados;
   - A atualização sistemática a que deverão estar sujeitos;
   - A avaliação da replicabilidade de soluções desenvolvidas;
   - O grau de supervisão humana sobre as decisões;
   - A capacidade para prever eventos raros;
   - A previsão e mensuração do impacto social resultante da sua aplicação;
   - As certificações necessárias;
   - Os indicadores de fiabilidade;
   - O impacto monetário nos organismos públicos;
   - A identificação de métricas para avaliar o treino e monitorização;

- A garantia de padrões de excelência científica;
- O grau de abertura na perspetiva de serem auditados e serem detetados eventuais erros.

RECOMENDAÇÕES À GESTÃO DE RISCO

Embora o valor da IA seja inquestionável pelo seu potencial de gerar benefícios para a sociedade, a utilização destes sistemas pode também desencadear **resultados indesejáveis significativos** para os indivíduos, organizações e sociedade.

A transição de uma posição de **avaliação** retrospetiva para uma **preventiva** torna-se essencial. A compreensão dos **riscos**, quais as suas interdependências e causas subjacentes, permitirá a sua identificação e priorização. Por sua vez, a identificação de riscos ocultos, incompreendidos ou não identificados, otimizará a sua deteção nos modelos mesmo antes de serem implementados.

Por sua vez, é importante que os investigadores e engenheiros dedicados à inteligência artificial tenham presente os efeitos duplos do seu trabalho, de modo a conseguirem definir objetivos e orientar conscientemente o seu trabalho, e **identificar situações de risco e vulnerabilidade, desencadeando as ações necessárias à sua prevenção ou mitigação**.

Como recomendações prioritárias à gestão de risco de implementação de serviços de IA, indicam-se:

- Aprendizagem com a comunidade de cibersegurança – formar equipas de verificação formal, divulgar vulnerabilidades da IA, desenvolver ferramentas de segurança e hardware seguro.
- Exploração de diferentes modelos de abertura – criar mecanismos e modelos de avaliação de risco de pré-publicação em áreas técnicas de preocupação especial, de licenciamento de acesso central e de partilha, que promovam a segurança e proteção.
- Promoção de uma cultura de responsabilização - educar, instituir diretrizes e padrões éticos, regulamentos, normas e diretivas.
- Desenvolvimento de soluções tecnológicas e políticas – respostas legislativas e regulatórias de proteção da privacidade, uso coordenado de IA para segurança do bem público, monitorização de recursos relevantes.

AGÊNCIA PARA A MODERNIZAÇÃO ADMINISTRATIVA
tic.gov.pt
Cofinanciado por:
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA
Fundo Social Europeu

Rua de Santa Marta, 55
1150-294 Lisboa | Portugal
guia@ama.pt