GRUPO INDEPENDENTE

DE PERITOS DE ALTO NÍVEL SOBRE A INTELIGÊNCIA ARTIFICIAL

CRIADO PELA COMISSÃO EUROPEIA EM JUNHO DE 2018

ORIENTAÇÕES ÉTICAS PARA UMA IA DE CONFIANÇA

Orientações éticas PARA UMA IA DE CONFIANÇA

Grupo de peritos de alto nível sobre a inteligência artificial

O presente documento foi elaborado pelo grupo de peritos de alto nível sobre a inteligência artificial (GPAN IA). Os membros do GPAN IA nomeados no presente documento apoiam o quadro geral para uma inteligência artificial de confiança proposto nas presentes orientações, embora não concordem necessariamente com todas as afirmações contidas no documento.

A lista de avaliação de uma IA de confiança, apresentada no capítulo III do presente documento, será aplicada pelas partes interessadas numa fase-piloto, tendo em vista a recolha de observações sobre a sua aplicação na prática. No início de 2020, será apresentada à Comissão Europeia uma versão revista da lista de avaliação tendo em conta as observações recolhidas durante a fase-piloto.

O GPAN IA é um grupo de peritos independente criado pela Comissão Europeia em junho de 2018.

Contacto Nathalie Smuha — Coordenadora do GPAN IA
Correio eletrónico CNECT-HLG-AI@ec.europa.eu

Comissão Europeia
B-1049 Bruxelas

Documento publicado em X de abril de 2019.

Em 18 de dezembro de 2018, foi publicada uma primeira versão deste documento, sobre a qual se pronunciaram mais de 500 participantes na consulta pública subsequentemente realizada. Desejamos agradecer expressamente e de forma calorosa a todos pelas suas observações sobre a primeira versão do documento, as quais foram tidas em consideração na elaboração da presente versão revista.

Estão disponíveis em linha mais informações sobre o grupo de peritos de alto nível sobre a inteligência artificial (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

ÍNDICE

RESUMO

1) O objetivo das presentes orientações é promover uma IA de confiança. Uma IA de confiança tem **três componentes**, que devem ser observadas ao longo de todo o ciclo de vida do sistema: a) deve ser **Legal**, cumprindo toda a legislação e regulamentação aplicáveis; b) deve ser **Ética**, garantindo a observância de princípios e valores éticos; c) deve ser **Sólida**, tanto do ponto de vista técnico como do ponto de vista social, uma vez que, mesmo com boas intenções, os sistemas de IA podem causar danos não intencionais. Cada uma destas componentes é necessária, mas não suficiente, para alcançar uma IA de confiança. Idealmente, as três componentes funcionam em harmonia, sobrepondo-se na sua ação. Se, na prática, surgirem conflitos entre elas, a sociedade deve procurar harmonizá-las.

2) As presentes orientações estabelecem um **quadro para alcançar uma IA de confiança**. O quadro não se ocupa explicitamente da primeira componente da IA de confiança (a IA legal)[1]. Ao invés, procura dar indicações sobre a forma de promover e assegurar uma IA ética e sólida (segunda e terceira componentes). Destinadas a todas as partes interessadas, estas orientações pretendem ser mais do que uma simples lista de princípios éticos, dando indicações sobre a forma de operacionalizar tais princípios em sistemas sociotécnicos. São facultadas orientações em três níveis de abstração, do nível mais abstrato, no capítulo I, até ao nível mais concreto, no capítulo III, terminando com exemplos de oportunidades e preocupações críticas suscitadas pelos sistemas de IA.

   I. Com base numa abordagem baseada nos direitos fundamentais, o capítulo I identifica os **princípios éticos** e respetivos valores que têm de ser respeitados durante o desenvolvimento, a implantação e a utilização dos sistemas de IA.

   Orientações fundamentais extraídas do capítulo I:

   - ✓ Desenvolver, implantar e utilizar os sistemas de IA de uma forma consentânea com os princípios éticos de: *respeito da autonomia humana, prevenção de danos, equidade e explicabilidade.* Reconhecer e procurar ultrapassar eventuais conflitos entre estes princípios.
   - ✓ Prestar especial atenção a situações que envolvam grupos mais vulneráveis, tais como crianças, pessoas com deficiência e outros grupos historicamente desfavorecidos ou em risco de exclusão, e a situações caracterizadas por assimetrias de poder ou de informação, como, por exemplo, entre empregadores e trabalhadores ou entre empresas e consumidores[2].
   - ✓ Reconhecer e ter presente que, embora tragam importantes benefícios para os indivíduos e a sociedade, os sistemas de IA apresentam também alguns riscos e são suscetíveis de ter um impacto negativo, incluindo impactos que podem ser difíceis de prever, identificar ou medir (p. ex., na democracia, no Estado de direito e na justiça distributiva, ou na própria mente humana). Adotar medidas adequadas para atenuar estes riscos quando necessário e proporcionalmente à dimensão do risco.

   II. Com base no capítulo I, o capítulo II fornece orientações sobre a forma de alcançar uma IA de confiança, enumerando **sete requisitos** que os sistemas de IA devem cumprir. Na sua aplicação, podem ser utilizados métodos técnicos e não técnicos.

[1] Todas as declarações normativas constantes do presente documento visam refletir uma orientação no sentido da concretização das segunda e terceira componentes de uma IA de confiança (a IA ética e sólida). Por conseguinte, as declarações não se destinam a fornecer aconselhamento jurídico ou orientações sobre o cumprimento da legislação aplicável, embora se reconheça que muitas destas declarações já estão, em alguma medida, refletidas na legislação existente. A este respeito, ver o ponto 21 e seguintes.

[2] Ver artigos 24.º a 27.º da Carta dos Direitos Fundamentais da União Europeia (Carta da UE), relativos aos direitos das crianças e dos idosos, à integração das pessoas com deficiência e aos direitos dos trabalhadores. Ver também o artigo 38.º relativo à defesa dos consumidores.

Orientações fundamentais extraídas do capítulo II:

- ✓ Assegurar que o desenvolvimento, a implantação e a utilização de sistemas de IA satisfazem os requisitos para uma IA de confiança: 1) ação e supervisão humanas; 2) solidez técnica e segurança; 3) privacidade e governação dos dados; 4) transparência; 5) diversidade, não discriminação e equidade; 6) bem-estar ambiental e societal; 7) responsabilização.
- ✓ Ponderar métodos técnicos e não técnicos para assegurar a aplicação desses requisitos.
- ✓ Promover a investigação e a inovação para ajudar a avaliar os sistemas de IA e a melhorar o cumprimento dos requisitos; divulgar os resultados e as questões em aberto junto do público em geral e formar sistematicamente uma nova geração de peritos em ética associada à IA.
- ✓ Comunicar, de forma clara e proativa, informações às partes interessadas sobre as capacidades e as limitações do sistema de IA, permitindo-lhes criar expectativas realistas, e sobre a forma como os requisitos são aplicados. Ser transparente sobre o facto de estarem a lidar com um sistema de IA.
- ✓ Facilitar a rastreabilidade e a auditabilidade dos sistemas de IA, sobretudo em contextos ou situações críticos.
- ✓ Envolver as partes interessadas em todo o ciclo de vida do sistema de IA. Promover a formação e a educação para que todas as partes interessadas tenham conhecimento e recebam formação em matéria de IA de confiança.
- ✓ Estar ciente de que podem existir conflitos fundamentais entre diferentes princípios e requisitos. Identificar, avaliar, documentar e comunicar continuamente essas soluções de compromisso.

III. O capítulo III apresenta uma lista concreta e não exaustiva de avaliação de uma IA de confiança, destinada a operacionalizar os requisitos enunciados no capítulo II. Esta **lista de avaliação** terá de ser adaptada ao caso de utilização específico do sistema de IA[3].

Orientações fundamentais extraídas do capítulo III:

- ✓ Adotar uma lista de avaliação para uma IA de confiança aquando do desenvolvimento, da implantação ou da utilização de sistemas de IA, e adaptá-la ao caso de utilização específico a que o sistema está a ser aplicado.
- ✓ Importa ter em mente que essa lista de avaliação nunca será exaustiva. Assegurar uma IA de confiança não se resume a um exercício de preenchimento de formulários; trata-se, sim, de um processo contínuo de identificação e aplicação de requisitos, de avaliação de soluções e de garantia de melhores resultados ao longo do ciclo de vida do sistema de IA, e de envolvimento das partes interessadas neste processo.

3) Uma secção final do documento visa concretizar algumas das questões abordadas no quadro, apresentando exemplos de oportunidades benéficas, que devem ser aproveitadas, e de preocupações críticas suscitadas pelos sistemas de IA, que devem ser cuidadosamente tomadas em consideração.

4) Embora as presentes orientações se destinem a dar indicações sobre as aplicações de IA em geral, criando uma base transversal para alcançar uma IA de confiança, situações diferentes suscitam diferentes desafios. Por conseguinte, é necessário analisar se, além deste quadro transversal, é necessária uma abordagem setorial, atendendo à especificidade contextual dos sistemas de IA.

5) As presentes orientações não se destinam a substituir qualquer forma de decisão política ou de regulamentação, atual ou futura, nem se destinam a dissuadir a sua introdução. Devem ser encaradas como um documento dinâmico, a rever e atualizar ao longo do tempo, para garantir que continua a ser relevante à

[3] De acordo com o âmbito do quadro definido no ponto 2, esta lista de avaliação não fornece quaisquer conselhos sobre o cumprimento da legislação (IA legal), limitando-se a facultar orientações sobre a concretização das segunda e terceira componentes de uma IA de confiança (IA ética e sólida).

medida que a tecnologia, os nossos ambientes sociais e os nossos conhecimentos evoluem. O presente documento é concebido como um ponto de partida para o debate sobre o tema «Uma inteligência artificial de confiança para a Europa»[4]. As orientações visam igualmente promover a investigação, a reflexão e o debate, para lá da Europa, sobre um quadro ético para os sistemas de IA a nível mundial.

[4] Este ideal destina-se a ser aplicado aos sistemas de IA desenvolvidos, implantados e utilizados nos Estados-Membros da UE, bem como aos sistemas desenvolvidos ou produzidos noutros países, mas implantados e utilizados na UE. No âmbito do presente documento, o termo «Europa» refere-se aos Estados-Membros da UE. No entanto, as presentes orientações aspiram também a ser relevantes fora da UE. A este respeito, importa salientar ainda que a Noruega e a Suíça fazem parte do Plano Coordenado para a Inteligência Artificial, acordado e publicado em dezembro de 2018 pela Comissão e pelos Estados-Membros.

A. INTRODUÇÃO

6) Nas suas comunicações de 25 de abril e 7 de dezembro de 2018, a Comissão Europeia (Comissão) apresentou a sua visão para a inteligência artificial (IA), que apoia uma «IA ética, segura e inovadora desenvolvida na Europa»[5]. A visão da Comissão está assente em três pilares: i) aumentar os investimentos públicos e privados na IA para reforçar a sua adoção, ii) preparar as mudanças socioeconómicas, iii) garantir um quadro ético e jurídico apropriado para reforçar os valores europeus.

7) Para apoiar a concretização desta visão, a Comissão criou o grupo de peritos de alto nível sobre a inteligência artificial (GPAN IA), um grupo independente incumbido da elaboração de dois documentos: 1) orientações éticas no domínio da IA; 2) recomendações políticas e de investimento.

8) O presente documento contém as orientações éticas no domínio da IA, as quais foram revistas na sequência de uma nova deliberação do grupo à luz das observações recebidas durante a consulta pública sobre o projeto publicado em 18 de dezembro de 2018. Tem ainda por base os trabalhos do Grupo Europeu de Ética para as Ciências e as Novas Tecnologias[6] e é inspirado noutros esforços similares[7].

9) Ao longo dos últimos meses, os 52 membros do grupo reuniram-se, debateram e interagiram, empenhados em concretizar a divisa europeia: unida na diversidade. Acreditamos que a IA tem potencial para transformar significativamente a sociedade. A IA não é um fim em si, mas antes um meio promissor para aumentar o desenvolvimento humano, reforçando, consequentemente, o bem-estar individual e societal e o bem comum, além de promover o progresso e a inovação. Os sistemas de IA podem contribuir, em particular, para facilitar a realização dos Objetivos de Desenvolvimento Sustentável da ONU, designadamente a promoção da igualdade de género e o combate às alterações climáticas, a racionalização da utilização dos recursos naturais, a melhoria da saúde, da mobilidade e dos processos de produção, bem como para apoiar a monitorização dos progressos alcançados com base em indicadores de sustentabilidade e coesão social.

10) Para este fim, os sistemas de IA[8] têm de estar **centrados no ser humano** e assentar no compromisso de serem utilizados ao serviço da humanidade e do bem comum, com o objetivo de melhorar o bem-estar e a liberdade dos seres humanos. Embora ofereçam grandes oportunidades, os sistemas de IA também apresentam certos riscos, que devem ser geridos de forma adequada e proporcionada. Temos agora uma importante janela de oportunidade para moldar o seu desenvolvimento. Queremos garantir que podemos confiar nos ambientes sociotécnicos em que eles estão incorporados e que os produtores dos sistemas de IA obtêm uma vantagem competitiva ao incorporarem uma IA de confiança nos seus produtos e serviços. Para que tal aconteça, há que procurar **maximizar os benefícios dos sistemas de IA, prevenindo e minimizando** simultaneamente **os seus riscos.**

11) Num contexto de rápida evolução tecnológica, consideramos essencial que a confiança continue a ser o elemento aglutinador das sociedades, das comunidades, das economias e do desenvolvimento sustentável. Por conseguinte, consideramos a **IA de confiança como a nossa ambição fundamental,** uma vez que os seres humanos e as suas comunidades só poderão confiar no desenvolvimento da tecnologia e nas suas aplicações se existir um quadro claro e abrangente que garanta a sua fiabilidade.

12) É este o caminho que acreditamos que a Europa deve seguir para afirmar a sua posição de liderança e promoção de tecnologias éticas e inovadoras. É através de uma IA de confiança que nós, cidadãos europeus,

---

[5] COM(2018) 237 e COM(2018) 795. Note-se que a expressão «*made in Europe*» (desenvolvida na Europa) é utilizada ao longo de toda a comunicação da Comissão. As presentes orientações visam, todavia, abranger não só os sistemas de IA desenvolvidos na Europa, mas também os desenvolvidos em países terceiros e implantados ou utilizados na Europa. Ao longo deste documento, pretende-se, assim, promover uma IA de confiança «para» a Europa.

[6] O Grupo Europeu de Ética para as Ciências e as Novas Tecnologias (EGE) é um grupo consultivo da Comissão.

[7] Ver ponto 3.3. do documento COM(2018) 237.

[8] No glossário incluído no final do presente documento, apresenta-se uma definição dos sistemas de IA que foi utilizada para efeitos da sua elaboração. Essa definição é aprofundada num documento específico elaborado pelo GPAN IA, que acompanha as presentes orientações, intitulado «Uma definição de IA: Principais capacidades e disciplinas científicas».

procuraremos colher os benefícios dessa tecnologia de forma consentânea com os nossos valores fundamentais de respeito dos direitos humanos, da democracia e do Estado de direito.

IA de confiança

13) A fiabilidade é uma condição prévia essencial para as pessoas e sociedades desenvolverem, implantarem e utilizarem os sistemas de IA. A impossibilidade de demonstrar que os sistemas de IA — e os seres humanos por detrás destes — são dignos de confiança poderá ter consequências indesejáveis, prejudicando a aceitação desses sistemas e impedindo a concretização dos vastos benefícios sociais e económicos que eles poderiam proporcionar. A nossa estratégia para ajudar a Europa a materializar esses benefícios consiste em utilizar a ética como um pilar fundamental para garantir e desenvolver uma IA de confiança.

14) A confiança no desenvolvimento, na implantação e na utilização dos sistemas de IA diz respeito não só às propriedades inerentes à tecnologia, mas também às qualidades dos sistemas sociotécnicos que envolvem aplicações de IA[9]. À semelhança das questões de (perda de) confiança na segurança da aviação, da energia nuclear ou dos alimentos, não são apenas as componentes do sistema de IA, mas o próprio sistema no seu contexto global, que podem, ou não, gerar confiança. Por conseguinte, os esforços para promover uma IA de confiança não só devem visar a fiabilidade do próprio sistema de IA, mas exigem uma abordagem holística e sistémica que abranja a fiabilidade de todos os intervenientes e processos que fazem parte do contexto sociotécnico do sistema ao longo do seu ciclo de vida.

15) Uma IA de confiança tem **três componentes**, que devem ser observadas ao longo de todo o ciclo de vida do sistema:

1. Deve ser **Legal**, garantindo o respeito de toda a legislação e regulamentação aplicáveis;
2. Deve ser **Ética**, garantindo a observância de princípios e valores éticos; e
3. Deve ser **Sólida**, tanto do ponto de vista técnico como do ponto de vista social, uma vez que, mesmo com boas intenções, os sistemas de IA podem causar danos não intencionais.

16) Cada uma destas três componentes é necessária, mas não suficiente, por si só, para alcançar uma IA de confiança[10]. O ideal é que as três funcionem em harmonia e se sobreponham na sua ação. Porém, na prática podem existir conflitos entre estes elementos (p. ex., o âmbito e o conteúdo da legislação em vigor podem não se coadunar, por vezes, com as normas éticas). Temos a responsabilidade individual e coletiva, enquanto sociedade, de procurar garantir que as três componentes contribuem para assegurar uma IA de confiança[11].

17) É fundamental que exista uma abordagem assente na confiança para permitir uma «competitividade responsável», estabelecendo a base para que todos os afetados pelos sistemas de IA possam confiar que a conceção, o desenvolvimento e a utilização desses sistemas são legais, éticos e sólidos. As presentes orientações destinam-se a promover uma inovação responsável e sustentável no domínio da IA a nível europeu. O seu intuito é fazer da ética um pilar fundamental do desenvolvimento de uma estratégia original para a IA, que vise beneficiar, capacitar e proteger tanto o desenvolvimento individual dos seres humanos como o bem comum da sociedade. Acreditamos que, desta forma, a Europa se poderá posicionar como líder mundial numa IA inovadora e digna da nossa confiança individual e coletiva. Só garantindo a fiabilidade dos sistemas de IA, os europeus conseguirão tirar pleno proveito dos seus benefícios, com a certeza de que existem medidas de salvaguarda contra os seus eventuais riscos.

18) Nem a utilização dos sistemas de IA nem o seu impacto conhecem fronteiras nacionais. Por conseguinte, são necessárias soluções a nível mundial para as oportunidades e os desafios globais resultantes da IA.

---

[9] Estes sistemas incluem seres humanos, intervenientes estatais, empresas, infraestruturas, *software*, protocolos, normas, governação, legislação em vigor, mecanismos de supervisão, estruturas de incentivo, procedimentos de auditoria, comunicação de melhores práticas e outros elementos.

[10] Isto não exclui a possibilidade de outras condições poderem ser (ou tornar-se) necessárias.

[11] Tal significa também que o corpo legislativo ou os decisores políticos podem ter de analisar a adequação da legislação em vigor, caso esta não se coadune com os princípios éticos.

Incentivamos, assim, todas as partes interessadas a trabalharem em prol da criação de um quadro mundial para uma IA de confiança, estabelecendo um consenso internacional, ao mesmo tempo que promovem e defendem a nossa abordagem baseada nos direitos fundamentais.

Audiência e âmbito

19) As presentes orientações são destinadas a todas as partes interessadas no domínio da IA que concebem, desenvolvem, implantam, aplicam, utilizam ou são afetadas pela IA, incluindo, mas não exclusivamente, empresas, organizações, investigadores, serviços públicos, organismos governamentais, instituições, organizações da sociedade civil, pessoas singulares, trabalhadores e consumidores. As partes interessadas empenhadas em alcançar uma IA de confiança podem decidir utilizar voluntariamente as presentes orientações como um método para operacionalizar o seu compromisso, recorrendo, em particular, à lista de avaliação prática constante do capítulo III nos seus processos de desenvolvimento e implantação de sistemas de IA. Esta lista de avaliação também pode complementar — e, logo, ser incorporada em — processos de avaliação existentes.

20) As orientações visam dar indicações sobre as aplicações de IA em geral, criando uma base transversal para alcançar uma IA de confiança. Contudo, **situações diferentes suscitam desafios diferentes.** Os sistemas de IA que recomendam músicas não levantam as mesmas preocupações éticas que os sistemas de IA que propõem tratamentos médicos de risco. Do mesmo modo, os sistemas de IA utilizados no contexto de relações empresas-consumidores, empresas-empresas, empregadores-trabalhadores e setor público-cidadãos ou, de um modo mais geral, em diferentes setores ou casos de utilização, apresentam oportunidades e desafios diferentes. Dada a especificidade contextual dos sistemas de IA, reconhece-se, portanto, que a utilização das presentes orientações tem de ser adaptada a cada aplicação de IA específica. Além disso, importa analisar se é necessário adotar uma abordagem setorial adicional, para complementar o quadro transversal mais geral proposto no presente documento.

Para compreender melhor como podem as presentes orientações ser aplicadas a nível transversal e que questões exigem uma abordagem setorial, convidamos todos os interessados a testarem a lista de avaliação de uma IA de confiança (capítulo III), que operacionaliza esse quadro, e a transmitirem-nos as suas observações sobre a experiência. Com base nas observações recolhidas no decurso desta fase-piloto, a lista de avaliação apresentada nas presentes orientações será revista, até ao início de 2020. A fase-piloto será lançada no verão de 2019 e prolongar-se-á até ao final do ano. Todas as partes interessadas poderão participar, manifestando o seu interesse através da Aliança Europeia para a IA.

B. QUADRO PARA UMA IA DE CONFIANÇA

21) As presentes orientações sistematizam um quadro para alcançar uma IA de confiança baseada nos direitos fundamentais consagrados na Carta dos Direitos Fundamentais da União Europeia (Carta da UE) e no direito internacional relevante em matéria de direitos humanos. Nos parágrafos seguintes, abordamos de forma sucinta as três componentes de uma IA de confiança.

IA legal

22) Os sistemas de IA não funcionam num mundo à margem da lei. Estão já em vigor várias regras juridicamente vinculativas a nível europeu, nacional e internacional, que são aplicáveis ou relevantes para o desenvolvimento, a implantação e a utilização de sistemas de IA. Entre as fontes jurídicas relevantes figuram, mas não exclusivamente, o direito primário da UE (os Tratados da União Europeia e a sua Carta dos Direitos Fundamentais), o direito derivado da UE (como o Regulamento Geral sobre a Proteção de Dados, as diretivas relativas à antidiscriminação, a Diretiva Máquinas, a Diretiva Produtos Defeituosos, o Regulamento Livre Fluxo

de Dados Não Pessoais e as diretivas nos domínios da defesa do consumidor e da segurança e saúde no trabalho), bem como os tratados da ONU em matéria de direitos humanos e as convenções do Conselho da Europa (como a Convenção Europeia dos Direitos Humanos) e muitas leis dos Estados-Membros da UE. Além das regras aplicáveis a nível transversal, existem várias regras setoriais para determinadas aplicações de IA (p. ex., o Regulamento Dispositivos Médicos no setor da saúde).

23) A legislação estabelece obrigações positivas e negativas, ou seja, deve ser interpretada não só à luz *do que não* pode ser feito, mas também do que *deve ser* feito. A legislação não só proíbe certas ações como também permite outras. Saliente-se, a este respeito, que a Carta da UE contém artigos sobre a «liberdade de empresa» e a «liberdade das artes e das ciências», em paralelo com artigos relativos a domínios que nos são mais familiares quando procuramos garantir a fiabilidade da IA, tais como a proteção de dados e a não discriminação.

24) As orientações não tratam explicitamente da primeira componente de uma IA de confiança (a IA legal), visando antes dar indicações sobre a forma de promover e garantir a segunda e a terceira componentes (IA ética e sólida). Embora, muitas vezes, as duas últimas já se encontrem, até certo ponto, refletidas na legislação em vigor, a sua plena realização pode ultrapassar as obrigações legais existentes.

25) Nada no presente documento deve ser entendido ou interpretado como uma forma de aconselhamento jurídico ou de orientação sobre o modo de cumprir normas e requisitos legais aplicáveis em vigor. Nada no presente documento cria direitos legais nem impõe obrigações legais em relação a terceiros. Todavia, recorde-se que todas as pessoas, singulares ou coletivas, têm o dever de cumprir a legislação — tanto a que é atualmente aplicável como a que será adotada no futuro de acordo com o desenvolvimento da IA. As presentes orientações partem do pressuposto de que **todos os direitos e obrigações legais aplicáveis aos processos e atividades implicados no desenvolvimento, na implantação e na utilização da IA continuam a ser de aplicação obrigatória e têm de ser devidamente respeitados.**

IA ética

26) Para alcançar uma IA de confiança é necessário cumprir a legislação, mas não só, pois esta é apenas uma das suas três componentes. A legislação nem sempre acompanha a rapidez da evolução tecnológica e, por vezes, pode estar desfasada das normas éticas ou não se adequar, pura e simplesmente, ao tratamento de certas questões. Por conseguinte, para os sistemas de IA serem confiáveis, devem também ser éticos e estar em harmonia com normas éticas.

IA sólida

27) Mesmo que se garanta que os sistemas de IA se destinam a um fim ético, as pessoas e as sociedades também devem estar confiantes de que eles não causarão danos não intencionais. Tais sistemas devem funcionar de forma segura e fiável, e devem prever-se salvaguardas para evitar impactos negativos não intencionais. Por conseguinte, é importante garantir a solidez dos sistemas de IA, tanto do ponto de vista técnico (assegurando a solidez técnica do sistema exigida em determinado contexto, tal como o domínio de aplicação ou a fase do ciclo de vida), como do ponto de vista social (tendo devidamente em conta o contexto e o ambiente em que o sistema opera). A IA ética e a IA sólida estão, assim, estreitamente interligadas e complementam-se entre si. Os princípios enunciados no capítulo I, e os requisitos correspondentes referidos no capítulo II, dizem respeito a ambas as componentes.

O quadro

28) No presente documento, as orientações são apresentadas em três níveis de abstração, começando pelo nível mais abstrato no capítulo I, e terminando no nível mais concreto no capítulo III:

I) Bases de uma IA de confiança. O capítulo I enuncia as bases de uma IA de confiança, definindo a sua

abordagem baseada nos direitos fundamentais[12]. Identifica e descreve os princípios éticos que devem ser respeitados para assegurar uma IA ética e sólida.

II) Concretização de uma IA de confiança. O capítulo II traduz estes princípios éticos em sete requisitos que os sistemas de IA devem aplicar e cumprir ao longo de todo o seu ciclo de vida. Além disso, propõe métodos técnicos e não técnicos, que podem ser utilizados na sua aplicação.

III) Avaliação de uma IA de confiança. Os profissionais no domínio da IA esperam orientações concretas. Por conseguinte, o capítulo III estabelece uma lista preliminar e não exaustiva de avaliação de uma IA de confiança para operacionalizar os requisitos do capítulo II. Esta avaliação deve ser adaptada à aplicação específica do sistema.

29) A secção final do documento apresenta as oportunidades benéficas e as preocupações críticas suscitadas pelos sistemas de IA que devem ser tomadas em consideração, e sobre as quais gostaríamos de estimular um debate mais aprofundado.

30) A estrutura das orientações é ilustrada na *figura 1* infra.

[12] Os direitos fundamentais constituem os alicerces do direito internacional e do direito da UE em matéria de direitos humanos e estão subjacentes aos direitos suscetíveis de proteção judicial garantidos pelos Tratados da UE e pela Carta dos Direitos Fundamentais da União Europeia. O cumprimento dos direitos fundamentais, sendo juridicamente vinculativo, está abrangido pela primeira componente de uma IA de confiança, a «IA legal». No entanto, pode considerar-se que os direitos fundamentais também refletem direitos especiais de caráter moral que pertencem a todos os indivíduos enquanto seres humanos, independentemente do seu estatuto juridicamente vinculativo. Nesse sentido, fazem igualmente parte da segunda componente de uma IA de confiança, a «IA ética».

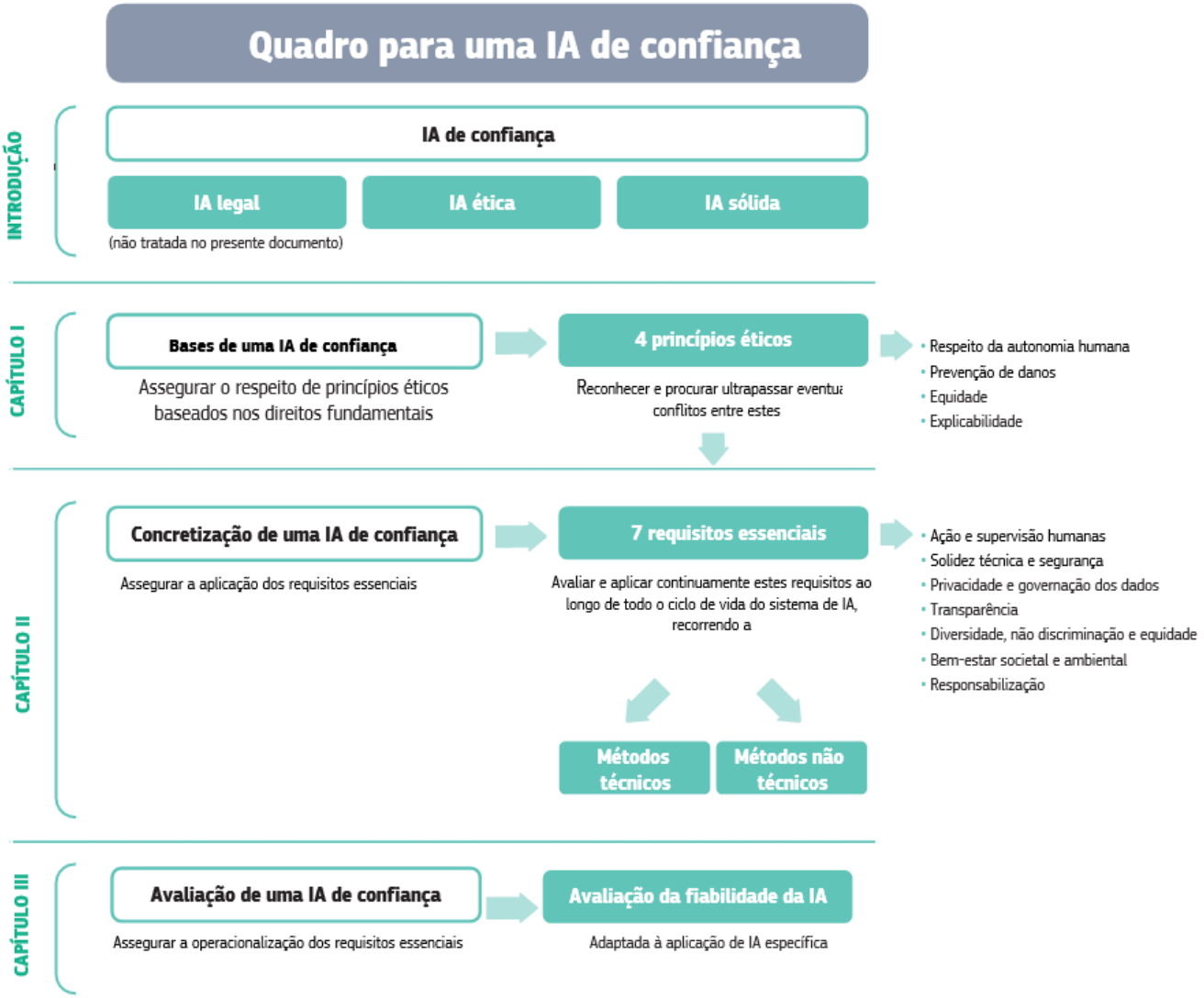

Figura 1: As orientações enquanto quadro para uma IA de confiança

I. Capítulo I: Bases de uma IA de confiança

31) No presente capítulo, definem-se as bases de uma IA de confiança, assentes nos direitos fundamentais e refletidas por quatro princípios éticos, que devem ser respeitados para garantir uma IA ética e sólida. Este capítulo é fortemente inspirado pelo domínio da ética.

32) A ética da IA é um subdomínio da ética aplicada que incide nas questões éticas suscitadas pelo desenvolvimento, pela implantação e pela utilização da inteligência artificial. A sua preocupação principal é identificar a forma como a IA pode melhorar ou suscitar preocupações para a vida das pessoas, quer em termos de qualidade de vida, quer de autonomia e liberdade humana necessárias para uma sociedade democrática.

33) A reflexão ética sobre a tecnologia de IA pode ter múltiplas finalidades. Em primeiro lugar, pode estimular a reflexão sobre a necessidade de proteger as pessoas e os grupos ao nível mais básico. Em segundo lugar, pode estimular novos tipos de inovações que procurem promover valores éticos, como os que contribuem para realizar os Objetivos de Desenvolvimento Sustentável da ONU[13], que estão firmemente incorporados na próxima Agenda 2030 da UE[14]. Embora o presente documento se ocupe principalmente da primeira finalidade referida, a importância que a ética poderá ter na segunda finalidade não deve ser subestimada. Uma IA de confiança pode melhorar o desenvolvimento individual e o bem-estar coletivo mediante a geração de prosperidade, a criação de valor e a maximização da riqueza. Pode contribuir para alcançar uma sociedade justa, ajudando a aumentar a saúde e o bem-estar dos cidadãos de forma a fomentar a igualdade na distribuição das oportunidades económicas, sociais e políticas.

34) Por conseguinte, importa compreendermos qual é a melhor forma de apoiar o desenvolvimento, a implantação e a utilização da IA para garantir que todos podem prosperar num mundo baseado na IA e construir um futuro melhor, mantendo simultaneamente a competitividade a nível mundial. Tal como acontece com qualquer tecnologia poderosa, a utilização de sistemas de IA na nossa sociedade suscita vários desafios éticos, nomeadamente em relação ao seu impacto nas pessoas e na sociedade, nas capacidades decisórias e na segurança. Para podermos recorrer cada vez mais à assistência dos sistemas de IA ou neles delegar decisões, temos de nos certificar de que esses sistemas produzem um impacto equitativo na vida das pessoas, de que são consentâneos com valores irredutíveis e capazes de agir de acordo com estes, e de que existem processos de responsabilização adequados para o garantir.

35) A Europa necessita de definir a visão normativa de um futuro imerso na IA que deseja realizar e de compreender, consequentemente, que conceito de IA deve ser estudado, desenvolvido, implantado e utilizado na Europa para concretizar essa visão. Com o presente documento, pretendemos contribuir para esse esforço mediante a introdução do conceito de IA de confiança, que acreditamos ser a forma correta de construir um futuro com inteligência artificial. Um futuro em que a democracia, o Estado de direito e os direitos fundamentais estão subjacentes aos sistemas de IA, e em que estes últimos melhoram e defendem continuamente a cultura democrática, também permitirá criar um ambiente em que a inovação e a competitividade responsável poderão prosperar.

36) Um código de ética para um domínio específico — por mais coerentes, desenvolvidas e pormenorizadas que as suas versões futuras possam ser — nunca poderá servir de substituto do próprio raciocínio ético, que deve permanecer sempre sensível a aspetos contextuais específicos, os quais não é possível apreender em orientações genéricas. Mais do que a definição de um conjunto de regras, garantir uma IA de confiança exige que construamos e mantenhamos uma cultura e uma mentalidade éticas através do debate público, da educação e da aprendizagem prática.

---

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_pt

14 https://sustainabledevelopment.un.org/?menu=1300.

1. Os direitos fundamentais como direitos de caráter moral e jurídico

37) Acreditamos numa abordagem à ética da IA baseada nos direitos fundamentais consagrados nos Tratados da UE[15], na Carta dos Direitos Fundamentais da União Europeia (Carta da UE) e no direito internacional em matéria de direitos humanos[16]. O respeito dos direitos fundamentais, num quadro de democracia e Estado de direito, proporciona as bases mais promissoras para identificar princípios e valores éticos abstratos que podem ser operacionalizados no contexto da IA.

38) Os Tratados da UE e a Carta da UE estabelecem uma série de direitos fundamentais, que os Estados-Membros e as instituições da UE são legalmente obrigados a respeitar quando aplicam o direito da União. Estes direitos são descritos na Carta da UE por referência à dignidade, às liberdades, à igualdade e solidariedade, aos direitos dos cidadãos e à justiça. Pode entender-se que a base comum que une estes direitos está alicerçada no respeito da dignidade humana — refletindo assim aquilo que designamos por «abordagem centrada no ser humano», na qual este goza de um estatuto moral único e inalienável de primazia nos domínios civil, político, económico e social[17].

39) Embora os direitos definidos na Carta da UE sejam juridicamente vinculativos[18], é importante reconhecer que os direitos fundamentais nem sempre proporcionam uma proteção jurídica global. No caso da Carta da UE, por exemplo, é importante sublinhar que o seu campo de aplicação está limitado aos domínios do direito da União. O direito internacional em matéria de direitos humanos e, em especial, a Convenção Europeia dos Direitos Humanos são juridicamente vinculativos para os Estados-Membros da UE, incluindo em domínios não abrangidos pelo direito da União. Simultaneamente, há que sublinhar que os direitos fundamentais também são conferidos a indivíduos e (até certo ponto) a grupos por força do seu estatuto moral como seres humanos, independentemente do seu valor jurídico. Entendidos como direitos suscetíveis de proteção judicial, os direitos fundamentais enquadram-se, por conseguinte, na primeira componente de uma IA de confiança (a IA legal), que garante o cumprimento da legislação. Entendidos como direitos universais, alicerçados no estatuto moral inerente aos seres humanos, estão também subjacentes à segunda componente de uma IA de confiança (a IA ética), respeitante a normas éticas que, embora não sejam necessariamente vinculativas em termos jurídicos, são cruciais para assegurar a fiabilidade. Uma vez que este documento não se destina a oferecer orientações sobre a primeira componente, para efeitos das presentes orientações não vinculativas, as referências aos direitos fundamentais estão relacionadas com a segunda componente.

2. Dos direitos fundamentais aos princípios éticos

2.1 Os direitos fundamentais como base de uma IA de confiança

40) Entre o conjunto abrangente de direitos indivisíveis definidos no direito internacional em matéria de direitos humanos, nos Tratados da UE e na Carta da UE, as categorias de direitos fundamentais abaixo referidas estão particularmente aptas a abarcar os sistemas de IA. Muitos destes direitos são, em determinadas circunstâncias, juridicamente vinculativos na UE, pelo que o cumprimento dos seus termos é legalmente obrigatório. Porém, mesmo depois de o cumprimento desses direitos fundamentais ter sido assegurado, a reflexão ética pode ajudar-nos a entender de que modo o desenvolvimento, a implantação e a utilização da IA

[15] A UE assenta num compromisso constitucional de proteger os direitos fundamentais e indivisíveis dos seres humanos, assegurar o respeito do Estado de direito, fomentar a liberdade democrática e promover o bem comum. Estes direitos estão refletidos nos artigos 2.º e 3.º do Tratado da União Europeia e na Carta dos Direitos Fundamentais da UE.

[16] Outros instrumentos jurídicos refletem e especificam melhor estes compromissos, como é o caso, por exemplo, da Carta Social Europeia do Conselho da Europa ou de legislação específica como o Regulamento Geral sobre a Proteção de Dados da UE.

[17] É de salientar que um compromisso no sentido de uma IA centrada no ser humano e alicerçada nos direitos fundamentais exige bases sociais e constitucionais coletivas em que a liberdade individual e o respeito da dignidade humana sejam simultaneamente possíveis em termos práticos e significativos, em vez de implicarem uma interpretação indevidamente individualista do ser humano.

[18] Nos termos do artigo 51.º da Carta, esta é aplicável às instituições e aos Estados-Membros da União quando apliquem o direito da União.

podem implicar os direitos fundamentais e os valores que lhes estão subjacentes, bem como a fornecer orientações mais pormenorizadas quando procurarmos identificar o que *devemos* fazer e não aquilo que *podemos* (atualmente) fazer com a tecnologia.

41) ***Respeito da dignidade humana.*** A dignidade humana engloba a ideia de que todos os seres humanos possuem um «valor intrínseco», que nunca deverá ser diminuído, posto em causa ou reprimido por outras pessoas — nem por novas tecnologias como os sistemas de IA[19]. No contexto da IA, o respeito da dignidade humana implica que todas as pessoas sejam tratadas com o respeito que lhes é devido enquanto *sujeitos* morais e não como meros *objetos* suscetíveis de serem examinados, triados, classificados, arregimentados, condicionados ou manipulados. Por conseguinte, os sistemas de IA devem ser desenvolvidos por forma a respeitar, servir e proteger a integridade física e mental dos seres humanos, o seu sentido de identidade pessoal e cultural e a satisfação das suas necessidades essenciais[20].

42) ***Liberdade do indivíduo.*** Os seres humanos devem ser livres de tomar decisões sobre as suas próprias vidas. Tal implica liberdade de intrusão soberana, mas também exige a intervenção de organizações governamentais e não governamentais para garantir que os indivíduos ou as pessoas em risco de exclusão têm igual acesso aos benefícios e oportunidades da IA. Num contexto de IA, a liberdade do indivíduo exige a atenuação da coerção ilegítima (in)direta, das ameaças à autonomia mental e à saúde mental, da vigilância injustificada, do engano e da manipulação indevida. De facto, a liberdade do indivíduo implica um compromisso para permitir que as pessoas exerçam um controlo ainda maior sobre as suas vidas, incluindo (entre outros direitos) a proteção do direito de empresa, a liberdade das artes e das ciências, a liberdade de expressão, o direito ao respeito da vida privada e familiar e a liberdade de reunião e de associação.

43) ***Respeito da democracia, da justiça e do Estado de direito.*** Nas democracias constitucionais, todo o poder estatal deve ser legalmente autorizado e limitado pela lei. Os sistemas de IA devem servir para manter e favorecer os processos democráticos e respeitar a pluralidade de valores e escolhas de vida dos indivíduos. Não devem comprometer os processos democráticos, a deliberação humana ou os sistemas de votação democráticos. Os sistemas de IA devem incorporar igualmente um compromisso de assegurar que o seu modo de funcionamento não prejudica os compromissos fundamentais em que o Estado de direito se baseia, nem a legislação e a regulamentação obrigatórias, bem como para garantir o direito a um processo justo e a igualdade perante a lei.

44) ***Igualdade, não discriminação e solidariedade*** *— incluindo os direitos das pessoas em risco de exclusão.* Deve ser assegurado o respeito igualitário do valor moral e da dignidade de todos os seres humanos. Tal vai além da não discriminação, que tolera o estabelecimento de distinções entre situações diferentes com base em justificações objetivas. Num contexto de IA, a igualdade implica que as operações do sistema não podem gerar resultados injustamente tendenciosos (p. ex., os dados utilizados para treinar os sistemas de IA devem ser o mais inclusivos possível, representando diferentes grupos da população). Tal exige que as pessoas e os grupos potencialmente vulneráveis[21], tais como trabalhadores, mulheres, pessoas com deficiência, minorias étnicas, crianças, consumidores, ou outras pessoas em risco de exclusão, sejam devidamente respeitados.

45) ***Direitos dos cidadãos.*** Os cidadãos usufruem de um amplo conjunto de direitos, incluindo o direito de voto, o direito a uma boa administração ou o direito de acesso a documentos públicos e o direito de petição à administração. Os sistemas de IA apresentam um significativo potencial para melhorar a escala e a eficiência das administrações públicas no fornecimento de bens e serviços públicos à sociedade. Ao mesmo tempo, os direitos dos cidadãos também podem ser negativamente afetados pelas aplicações de IA, pelo devem ser protegidos. Quando o termo «direitos dos cidadãos» é aqui utilizado, o intuito não é negar ou negligenciar os

[19] C. McCrudden, «Human Dignity and Judicial Interpretation of Human Rights», EJIL, 19(4), 2008.

[20] Para uma compreensão do conceito de «dignidade humana» neste sentido, ver E. Hilgendorf, «Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity», em: D. Grimm, A. Kemmerer, C. Möllers (editores), Human Dignity in Context. Explorations of a Contested Concept, 2018, p. 325 e seguintes.

[21] O glossário contém uma descrição do termo tal como é utilizado ao longo do presente documento.

direitos dos nacionais de países terceiros e das pessoas em situação irregular (ou ilegal) na UE, que também têm direitos ao abrigo do direito internacional e, logo, no domínio da IA.

2.2 Princípios éticos no contexto dos sistemas de IA[22]

46) Muitas organizações públicas, privadas e da sociedade civil inspiraram-se nos direitos humanos para produzirem enquadramentos éticos para a IA[23]. Na UE, o Grupo Europeu de Ética para as Ciências e as Novas Tecnologias («EGE») propôs um conjunto de nove princípios básicos, com base nos valores fundamentais estabelecidos nos Tratados da UE e na Carta dos Direitos Fundamentais da União Europeia[24]. Continuamos a desenvolver esse trabalho, reconhecendo a maioria dos princípios até agora defendidos por vários grupos, clarificando simultaneamente os fins que todos os princípios procuram fomentar e apoiar. Estes princípios éticos podem inspirar instrumentos regulamentares novos e específicos, ajudar a interpretar os direitos fundamentais à medida que o nosso ambiente sociotécnico evolui ao longo do tempo e orientar a fundamentação do desenvolvimento, da utilização e da aplicação dos sistemas de IA, adaptando-se de forma dinâmica à evolução da própria sociedade.

47) Os sistemas de IA devem melhorar o bem-estar individual e coletivo. Esta secção enumera **quatro princípios éticos**, enraizados nos direitos fundamentais, que devem ser respeitados para assegurar que os sistemas de IA são desenvolvidos, implantados e utilizados de forma confiável. São especificados como **imperativos éticos**, que os profissionais no domínio da IA devem esforçar-se sempre por respeitar. Sem impor uma hierarquia, enumeramos a seguir os princípios de modo a refletir a ordem em que os direitos fundamentais nos quais se baseiam são apresentados na Carta da UE[25].

48) Estes são os princípios de:

i) Respeito da autonomia humana
ii) Prevenção de danos
iii) Equidade
iv) Explicabilidade

49) Muitos deles já estão, em grande medida, refletidos nos requisitos jurídicos existentes, que têm de ser obrigatoriamente cumpridos, pelo que estão também abrangidos pela «IA legal», que constitui a primeira componente de uma IA de confiança[26]. Todavia, como é acima referido, embora muitas obrigações legais reflitam princípios éticos, o respeito dos mesmos vai além do cumprimento formal da legislação existente[27].

- O princípio do respeito da autonomia humana

50) Os direitos fundamentais em que a UE se alicerça visam garantir o respeito da liberdade e da autonomia dos seres humanos. Os seres humanos que interajam com sistemas de IA devem poder manter uma

[22] Estes princípios também são aplicáveis ao desenvolvimento, à implantação e à utilização de outras tecnologias e, por conseguinte, não são específicos dos sistemas de IA. No texto seguinte, procurámos definir a sua relevância especificamente num contexto relacionado com a IA.

[23] O respeito dos direitos fundamentais também ajuda a limitar a insegurança regulamentar, dado que pode basear-se em décadas de prática de proteção dos direitos fundamentais na UE, proporcionando assim clareza, legibilidade e previsibilidade.

[24] Mais recentemente, o grupo de trabalho do AI4People realizou um inquérito sobre os princípios do EGE acima referidos, bem como sobre 36 outros princípios éticos apresentados até à data, e agrupou-os em quatro princípios gerais: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), «AI4People — An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations», Minds and Machines 28(4): pp. 689-707.

[25] O respeito da autonomia humana está fortemente associado ao direito à dignidade do ser humano e à liberdade (refletidos nos artigos 1.º a 6.º da Carta). A prevenção de danos está fortemente ligada à proteção da integridade física e mental (refletida no artigo 3.º). A equidade está estreitamente ligada aos direitos à não discriminação, à solidariedade e à justiça (refletidos nos artigos 21.º e seguintes). A explicabilidade e a responsabilidade estão estreitamente ligadas aos direitos relacionados com a justiça (tal como refletidos no artigo 47.º).

[26] Veja-se, por exemplo, o RGPD ou os regulamentos da UE em matéria de defesa do consumidor.

[27] Para saber mais sobre este tema, ver, por exemplo, L. Floridi, «Soft Ethics and the Governance of the Digital», Philosophy & Technology, março de 2018, volume 31, n.º 1, p. 1–8.

autodeterminação plena e efetiva sobre si próprios e participar no processo democrático. Os sistemas de IA não devem subordinar, coagir, enganar, manipular, condicionar ou arregimentar injustificadamente os seres humanos. Em vez disso, os sistemas de IA devem ser concebidos para aumentar, complementar e capacitar as competências cognitivas, sociais e culturais dos seres humanas. A distribuição de funções entre os seres humanos e os sistemas de IA devem seguir princípios de conceção centrados no ser humano e deixar uma oportunidade significativa para a escolha humana. Isto implica que se garanta a supervisão[28] e o controlo por parte de seres humanos sobre os processos de trabalho dos sistemas de IA. Estes sistemas também podem alterar radicalmente a esfera do trabalho, que deverá apoiar os seres humanos no ambiente de trabalho e visar a criação de um trabalho significativo.

- O princípio da prevenção de danos

51) Os sistemas de IA não devem causar danos ou agravá-los[29] nem afetar negativamente os seres humanos de qualquer outra forma[30]. Isto implica a proteção da dignidade, bem como da integridade mental e física, do ser humano. Os sistemas de IA e os ambientes em que operam devem ser seguros e protegidos. Devem ser tecnicamente sólidos e deve garantir-se que não estão abertos a utilizações malévolas. As pessoas vulneráveis devem receber maior atenção e ser incluídas no desenvolvimento e na implantação dos sistemas de IA. Há também que prestar especial atenção às situações em que os sistemas de IA podem causar ou agravar impactos negativos devido a assimetrias de poder ou de informação, nomeadamente entre empregadores e trabalhadores, empresas e consumidores ou governos e cidadãos. A prevenção dos danos implica também ter em consideração o ambiente natural e todos os seres vivos.

- O princípio da equidade

52) O desenvolvimento, a implantação e a utilização dos sistemas de IA devem ser equitativos. Embora reconheçamos que há muitas interpretações diferentes de equidade, consideramos que esta tem uma dimensão substantiva e processual. A dimensão substantiva implica um compromisso com: a garantia de uma distribuição equitativa e justa dos benefícios e dos custos, bem como de inexistência de enviesamentos injustos, discriminação e estigmatização contra pessoas e grupos. Se for possível evitar os enviesamentos, os sistemas de IA podem até aumentar a equidade societal. A igualdade de oportunidades em termos de acesso à educação, aos bens e serviços e à tecnologia deve ser igualmente promovida. Além disso, a utilização de sistemas de IA nunca deverá levar a que os utilizadores (finais) sejam iludidos ou prejudicados na sua liberdade de escolha. Além disso, a equidade implica que os profissionais no domínio da IA devem respeitar o princípio da proporcionalidade entre os meios e os fins, e analisar cuidadosamente a forma de equilibrar os interesses e objetivos em causa[31]. A dimensão processual da equidade implica uma possibilidade de contestar e procurar vias de recurso eficazes contra as decisões tomadas por sistemas de IA e pelos seres humanos que os utilizam[32]. Para o efeito, a entidade responsável pela decisão deve ser identificável e os processos decisórios explicáveis.

---

28 O conceito de supervisão humana é desenvolvido no ponto 65 infra.

29 Os danos podem ser individuais ou coletivos e incluir danos intangíveis para os ambientes sociais, culturais e políticos.

30 Isto abrange igualmente o modo de vida dos indivíduos e dos grupos sociais, evitando, por exemplo, os danos culturais.

31 Esta questão está relacionada com o princípio da proporcionalidade (refletido no aforismo de que não se deve «matar moscas com balas de canhão»). As medidas tomadas para atingir um fim (p. ex., as medidas de extração de dados aplicadas para concretizar a função de otimização da IA) devem ser limitadas ao estritamente necessário. Implica igualmente que, quando diversas medidas concorrem entre si para a consecução de um fim, deve dar-se preferência à que for menos contrária aos direitos fundamentais e às normas éticas (p. ex., os criadores de IA devem preferir sempre os dados do setor público aos dados pessoais). Pode fazer-se igualmente referência à proporcionalidade entre o utilizador e o implantador, tomando em consideração os direitos das empresas (incluindo os de propriedade intelectual e de confidencialidade), por um lado, e os direitos do utilizador, por outro lado.

32 Incluindo mediante a utilização do seu direito de associação e de filiação num sindicato, tal como previsto no artigo 12.º da Carta dos Direitos Fundamentais da União Europeia.

- O princípio da explicabilidade

53) A explicabilidade é crucial para criar e manter a confiança dos utilizadores nos sistemas de IA. Tal significa que os processos têm de ser transparentes, as capacidades e a finalidade dos sistemas de IA abertamente comunicadas e as decisões — tanto quanto possível — explicáveis aos que são por elas afetados de forma direta e indireta. Sem essas informações, não é possível contestar devidamente uma decisão. Nem sempre é possível explicar por que razão um modelo gerou determinado resultado ou decisão (e que combinação de fatores de entrada contribuiu para esse efeito). Estes casos são designados por algoritmos de «caixa negra» e exigem especial atenção. Nessas circunstâncias, podem ser necessárias outras medidas da explicabilidade (p. ex., a rastreabilidade, a auditabilidade e a comunicação transparente sobre as capacidades do sistema), desde que o sistema, no seu conjunto, respeite os direitos fundamentais. O grau de necessidade da explicabilidade depende em grande medida do contexto e da gravidade das consequências de um resultado errado ou inexato[33].

2.3 Conflitos entre os princípios

54) Podem surgir conflitos entre os princípios acima enunciados e não há uma solução rígida para os resolver. Em consonância com o compromisso fundamental da UE para com o empenhamento democrático, o direito a um processo justo e uma participação política aberta, devem ser estabelecidos métodos de deliberação responsável para fazer face a esses conflitos. Por exemplo, em vários domínios de aplicação, *o princípio da prevenção de danos* e *o princípio da autonomia humana* podem entrar em conflito. Veja-se o exemplo da utilização de sistemas de IA para a «previsão policial», que podem ajudar a reduzir a criminalidade, mas recorrendo a atividades de vigilância que colidem com a liberdade e a privacidade individuais. Além disso, os benefícios globais dos sistemas de IA devem ser substancialmente superiores aos riscos individuais previsíveis. Embora seja indubitável que estes princípios oferecem uma orientação no sentido de encontrar soluções, continuam a ser preceitos éticos abstratos. Por conseguinte, não se pode esperar que os profissionais no domínio da IA encontrem a solução adequada com base nos princípios acima referidos, mas devem abordar os dilemas éticos e as soluções de compromisso através de uma reflexão racional, baseada em factos e não na intuição ou numa apreciação aleatória. Podem existir, no entanto, situações em que não seja possível identificar quaisquer soluções de compromisso eticamente aceitáveis. Determinados direitos fundamentais e princípios com eles relacionados são absolutos e não podem ser objeto de ponderação (p. ex., a dignidade humana).

Orientações fundamentais extraídas do capítulo I:

- ✓ Desenvolver, implantar e utilizar os sistemas de IA de uma forma consentânea com os princípios éticos de: *respeito da autonomia humana, prevenção de danos, equidade e explicabilidade.* Reconhecer e procurar ultrapassar eventuais conflitos entre estes princípios.
- ✓ Prestar especial atenção a situações que envolvam grupos mais vulneráveis, tais como crianças, pessoas com deficiência e outros grupos historicamente desfavorecidos ou em risco de exclusão, e/ou a situações caracterizadas por assimetrias de poder ou de informação, como, por exemplo, entre empregadores e trabalhadores ou entre empresas e consumidores[34].
- ✓ Reconhecer e ter presente que, embora tenham potencial para trazer muitos benefícios substanciais para os indivíduos e a sociedade, algumas aplicações de IA também são suscetíveis de ter impactos negativos, incluindo alguns que podem ser difíceis de prever, identificar ou medir (p. ex., na democracia, no Estado

[33] Por exemplo, as recomendações de compra inexatas geradas por um sistema de IA podem suscitar poucas preocupações éticas, ao contrário dos sistemas de IA que avaliam se uma pessoa condenada por uma infração penal deve ou não beneficiar de liberdade condicional.

[34] Ver artigos 24.º a 27.º da Carta da UE, relativos aos direitos das crianças e dos idosos, à integração das pessoas com deficiência e aos direitos dos trabalhadores. Ver também o artigo 38.º relativo à defesa dos consumidores.

de direito e na justiça distributiva, ou na própria mente humana). Adotar medidas adequadas para atenuar estes riscos quando necessário e proporcionalmente à dimensão do risco.

II. Capítulo II: Concretização de uma IA de confiança

55) O presente capítulo fornece orientações sobre a aplicação e a concretização de uma IA de confiança, através de uma lista de sete requisitos a cumprir, com base nos princípios descritos no capítulo I. Além disso, apresentam-se métodos técnicos e não técnicos atualmente disponíveis para aplicar estes requisitos ao longo de todo o ciclo de vida do sistema de IA.

1. Requisitos de uma IA de confiança

56) Os princípios descritos no capítulo I devem ser traduzidos em requisitos concretos para alcançar uma IA de confiança. Estes requisitos são aplicáveis a diversas partes interessadas participantes no ciclo de vida dos sistemas de IA: criadores, implantadores e utilizadores finais, bem como a sociedade em geral. Designamos por criadores aqueles que investigam, concebem e/ou desenvolvem os sistemas de IA. Por implantadores, entendemos as organizações públicas ou privadas que utilizam os sistemas de IA nos seus processos empresariais e para oferecer produtos e serviços a outros. Os utilizadores finais são as pessoas que interagem de forma direta ou indireta com o sistema de IA. Por último, a sociedade em geral engloba todas as outras pessoas que são direta ou indiretamente afetadas pelos sistemas de IA.

57) As diferentes classes de partes interessadas desempenham papéis diferentes no que toca a garantir que os requisitos são cumpridos:

a. Os criadores devem adotar e aplicar os requisitos aos processos de conceção e desenvolvimento;
b. Os implantadores devem assegurar que os sistemas que utilizam e os produtos e serviços que oferecem cumprem os requisitos;
c. Os utilizadores finais e a sociedade em geral devem ser informados acerca destes requisitos e poder exigir que os mesmos sejam respeitados.

58) A lista de requisitos a seguir apresentada não é exaustiva[35]. Inclui aspetos sistémicos, individuais e societais:

1 Ação e supervisão humanas
Incluindo os direitos fundamentais, a ação humana e a supervisão humana

2 Solidez técnica e segurança
Incluindo a resiliência perante ataques e a segurança, os planos de recurso e a segurança geral, a exatidão, a fiabilidade e a reprodutibilidade

3 Privacidade e governação dos dados
Incluindo o respeito da privacidade, a qualidade e a integridade dos dados e o acesso aos dados

4 Transparência
Incluindo a rastreabilidade, a explicabilidade e a comunicação

5 Diversidade, não discriminação e equidade
Incluindo a prevenção de enviesamentos injustos, a acessibilidade e a conceção universal e a participação das partes interessadas

6 Bem-estar societal e ambiental
Incluindo a sustentabilidade e o respeito do ambiente, o impacto social, a sociedade e a democracia

[35] Sem impor uma hierarquia, enumeramos a seguir os princípios de modo a refletir a ordem em que os princípios e os direitos com que estão relacionados são apresentados na Carta da UE.

7 Responsabilização

Incluindo a auditabilidade, a minimização e a comunicação dos impactos negativos, as soluções de compromisso e as vias de recurso.

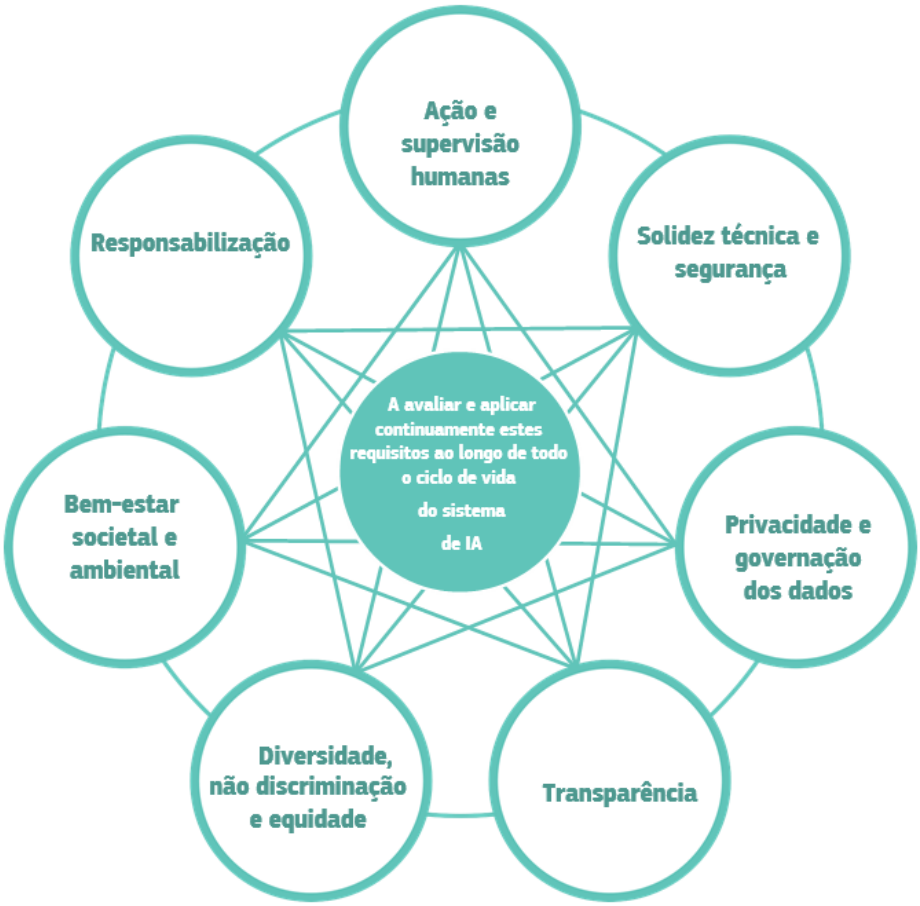

Figura 2: Interligação dos sete requisitos: todos têm igual importância e apoiam-se mutuamente, devendo ser aplicados e avaliados ao longo de todo o ciclo de vida de um sistema de IA

59) Embora todos requisitos sejam igualmente importantes, o contexto e os eventuais conflitos entre eles terão de ser tidos em conta quando forem aplicados nos diversos domínios e setores. A aplicação destes requisitos deve ocorrer ao longo de todo o ciclo de vida de um sistema de IA e depende da aplicação em causa. Embora os requisitos sejam, na sua maioria, aplicáveis a todos os sistemas de IA, é dada especial atenção aos que afetam direta ou indiretamente os indivíduos. Por conseguinte, no caso de algumas aplicações (p. ex., em contextos industriais), podem ter menos relevância.

60) Os requisitos acima referidos incluem elementos que, em alguns casos, já estão refletidos na legislação existente. Reiteramos que — em consonância com a primeira componente de uma IA de confiança — compete aos criadores e implantadores dos sistemas de IA garantir que cumprem as suas obrigações legais, tanto no que diz respeito às regras transversalmente aplicáveis como à regulamentação setorial.

61) Nos parágrafos seguintes, cada requisito é tratado de forma mais aprofundada.

1. <u>Ação e supervisão humanas</u>

62) Os sistemas de IA devem apoiar a autonomia e a tomada de decisões dos seres humanos, tal como prescrito

pelo princípio de *respeito da autonomia humana*. Isto exige que os sistemas de IA funcionem como facilitadores de uma sociedade democrática, próspera e equitativa, apoiando a ação do utilizador e a promoção dos direitos fundamentais, e que permitam a supervisão humana.

63) ***Direitos fundamentais***. À semelhança de muitas outras tecnologias, os sistemas de IA tanto podem favorecer como prejudicar o usufruto dos direitos fundamentais. Podem beneficiar as pessoas, por exemplo, ajudando-as a rastrear os seus dados pessoais ou aumentando o seu acesso à educação e apoiando, assim, o direito à mesma. No entanto, dado o alcance e a capacidade dos sistemas de IA, também podem afetar negativamente a defesa dos direitos fundamentais. Nas situações em que existem riscos como estes, deve realizar-se uma avaliação do impacto nos direitos fundamentais antes do desenvolvimento dos sistemas, a qual deve incluir uma avaliação da possibilidade de reduzir ou justificar esses riscos na medida do necessário numa sociedade democrática, a fim de respeitar os direitos e liberdades dos outros. Além disso, devem ser criados mecanismos para receber observações externas sobre os sistemas de IA que possam contrariar os direitos fundamentais.

64) ***Ação humana.*** Os utilizadores devem poder tomar decisões autónomas e fundamentadas a respeito dos sistemas de IA. Devem ser-lhes facultados conhecimentos e ferramentas para compreenderem e interagirem com os sistemas de IA a um nível satisfatório e, se possível, deve ser-lhes dada a possibilidade de eles próprios avaliarem ou contestarem o sistema. Os sistemas de IA devem ajudar os indivíduos a fazerem escolhas mais corretas e fundamentadas em conformidade com os seus objetivos. Por vezes, os sistemas de IA podem ser utilizados para moldar e influenciar o comportamento humano mediante mecanismos talvez difíceis de detetar por utilizarem processos subconscientes, incluindo várias formas desleais de manipulação, engano, arregimentação e condicionamento, todas elas suscetíveis de pôr em risco a autonomia individual. O princípio geral da autonomia do utilizador deve estar no centro da funcionalidade do sistema. Destaca-se, a este respeito, o direito a não ficar sujeito a nenhuma decisão tomada exclusivamente com base no tratamento automatizado quando esta produza efeitos na esfera jurídica dos utilizadores ou os afete significativamente de forma similar[36].

65) ***Supervisão humana***. A supervisão humana ajuda a garantir que um sistema de IA não põe em causa a autonomia humana nem produz outros efeitos negativos. A supervisão pode ser realizada mediante mecanismos de governação como as abordagens de intervenção humana (*human-in-the-loop* — HITL), de fiscalização humana (*human-on-the-loop* — HOTL), ou de controlo humano (*human-in-command* — HIC). A abordagem HITL refere-se à capacidade de intervenção humana em todos os ciclos de decisão do sistema, a qual, em muitos casos, não é possível nem desejável. A abordagem HOTL refere-se à capacidade de intervenção humana durante o ciclo de conceção do sistema e de acompanhamento do funcionamento do sistema. A abordagem HIC refere-se à capacidade de supervisionar toda a atividade do sistema de IA (incluindo o seu impacto económico, societal, jurídico e ético mais geral) e de decidir quando e como utilizar o sistema em qualquer situação específica. Tal pode incluir a decisão de não utilizar um sistema de IA numa determinada situação, de estabelecer níveis de apreciação humana durante a utilização do sistema, ou de assegurar a capacidade de anular uma decisão tomada por um sistema. Além disso, deve ser garantido que as autoridades públicas responsáveis pela aplicação da lei têm a possibilidade de exercer a supervisão em conformidade com o seu mandato. Podem ser necessários mecanismos de supervisão em graus variáveis para apoiar outras medidas de segurança e controlo, dependendo do domínio de aplicação e do potencial risco do sistema de IA. Não havendo alteração das demais condições, quanto menor for a supervisão que um ser humano pode exercer sobre um sistema de IA, maior será a necessidade de sujeitar o mesmo a amplos testes e a uma governação rigorosa.

[36] É de referir o artigo 22.º do RGPD, em que este direito já se encontra consagrado.

2. Solidez técnica e segurança

66) Uma componente crucial para que a IA de confiança se torne realidade é a solidez técnica, que está estreitamente ligada ao *princípio da prevenção de danos*. A solidez técnica exige que os sistemas de IA sejam desenvolvidos seguindo uma abordagem de prevenção dos riscos e de forma a que se comportem fiavelmente conforme o previsto, minimizando os danos não intencionais e inesperados, e prevenindo os danos inaceitáveis. Tal deverá também ser aplicado a eventuais alterações do ambiente em que operam ou à presença de outros agentes (humanos e artificiais) que possam interagir com o sistema de forma antagónica. Além disso, deve assegurar-se a integridade física e mental dos seres humanos.

67) ***Resiliência perante ataques e segurança***. Os sistemas de IA, à semelhança de todos os outros sistemas informáticos, devem ser protegidos contra vulnerabilidades que permitam a sua exploração por adversários, por exemplo, piratas informáticos. Os ataques podem ser dirigidos contra os dados (adulteração de dados — *data poisoning*), o modelo (fuga de modelos — *model leakage*) ou a infraestrutura subjacente, tanto de software como de hardware. Se um sistema de IA for atacado, por exemplo no caso dos ataques antagónicos, os dados e o comportamento do sistema podem ser alterados, fazendo com este tome decisões diferentes ou se desligue completamente. Os sistemas e os dados também podem ser corrompidos com intenção malévola ou pela exposição a situações inesperadas. Os processos de segurança insuficientes também podem dar origem a decisões erradas ou mesmo a danos físicos. Para os sistemas de IA serem considerados seguros[37], há que ter em conta as possíveis aplicações não intencionais da IA (p. ex., aplicações de dupla utilização) e o eventual abuso de um sistema de IA por intervenientes mal-intencionados, e devem ser tomadas medidas para os atenuar[38].

68) ***Plano de recurso e segurança geral***. Todos os sistemas de IA devem possuir salvaguardas que possibilitem um plano de recurso em caso de problemas. Tal pode implicar que os sistemas de IA mudem de um procedimento estatístico para um procedimento baseado em regras, ou que solicitem a intervenção de um operador humano antes de continuarem a sua ação[39]. Deve garantir-se que o sistema funcionará como previsto, sem causar danos aos seres humanos nem ao ambiente, o que também implica uma minimização das consequências não intencionais e dos erros. Além disso, devem adotar-se processos para clarificar e avaliar os potenciais riscos associados à utilização de sistemas de IA em diversos domínios de aplicação. O nível das medidas de segurança necessárias depende da magnitude do risco colocado por um sistema de IA, a qual depende por sua vez das capacidades do sistema. Quando seja possível prever riscos particularmente elevados no processo de desenvolvimento ou no próprio sistema, é crucial desenvolver e testar medidas de segurança de forma proativa.

69) ***Exatidão.*** A exatidão diz respeito à capacidade do sistema de IA para fazer apreciações corretas, por exemplo para classificar corretamente as informações nas categorias adequadas, ou à sua capacidade para formular previsões, recomendações ou decisões corretas com base em dados ou modelos. Um processo de desenvolvimento e avaliação explícito e bem formado pode apoiar, atenuar e corrigir riscos não intencionais decorrentes de previsões incorretas. Quando não for possível evitar previsões incorretas ocasionais, é importante que o sistema possa indicar a probabilidade de tais erros ocorrerem. Um elevado nível de exatidão é particularmente crucial em situações em que o sistema de IA afeta diretamente vidas humanas.

70) ***Fiabilidade e reprodutibilidade.*** É essencial que os resultados dos sistemas de IA possam ser reproduzidos e que sejam fiáveis. Um sistema de IA é considerado fiável se funcionar adequadamente com vários tipos de

---

37 Ver, por exemplo, as considerações formuladas no ponto 2.7 do Plano Coordenado para a Inteligência Artificial da União Europeia.

38 Para garantir a segurança dos sistemas de IA, pode ser indispensável conseguir desenvolver um círculo virtuoso na investigação e no desenvolvimento entre a compreensão dos ataques, a elaboração de medidas de proteção adequadas e a melhoria das metodologias de avaliação. Para o efeito, há que promover uma convergência entre a comunidade de IA e a comunidade de segurança. Além disso, compete a todos os intervenientes envolvidos criar normas transfronteiriças comuns de segurança e proteção, bem como estabelecer um ambiente de confiança mútua, promovendo a colaboração a nível internacional. Relativamente a medidas possíveis, ver Malicious Use of AI (Avin S., Brundage M., et. al., 2018).

39 Os cenários em que a intervenção humana não seria imediatamente possível devem ser igualmente tidos em conta.

dados de entrada e em várias situações. Tal é necessário para examinar pormenorizadamente um sistema de IA e prevenir danos não intencionais. A reprodutibilidade descreve se uma experiência de IA apresenta o mesmo comportamento quando repetida nas mesmas condições. Isto permite que os cientistas e os decisores políticos descrevam com exatidão o que os sistemas de IA fazem. Os ficheiros de replicação[40] podem facilitar o processo de experimentação e reprodução dos comportamentos.

3. Privacidade e governação dos dados

71) Estreitamente ligado ao *princípio de prevenção de danos* está o direito à privacidade, um direito fundamental que é particularmente afetado pelos sistemas de IA. A prevenção da ameaça à privacidade também exige uma governação adequada dos dados, que assegure a qualidade e a integridade dos dados utilizados, a sua relevância para o domínio em que os sistemas de IA serão implantados, os seus protocolos de acesso e a capacidade de tratar os dados de modo a proteger a privacidade.

72) ***Privacidade e proteção de dados***. Os sistemas de IA devem garantir a privacidade e a proteção de dados ao longo de todo o ciclo de vida de um sistema[41]. Tal inclui as informações inicialmente fornecidas pelo utilizador, bem como as informações produzidas sobre o utilizador ao longo da sua interação com o sistema (p. ex., os resultados gerados pelo sistema de IA para utilizadores específicos ou a forma como os utilizadores responderam a determinadas recomendações). Os registos digitais do comportamento humano podem permitir que os sistemas de IA infiram não só as preferências dos indivíduos, mas também a sua orientação sexual, a sua idade e as suas convicções religiosas ou políticas. Para que as pessoas possam confiar no processo de recolha de dados, deve ser garantido que os dados recolhidos a seu respeito não serão utilizados para as discriminar de forma ilegal ou injusta.

73) ***Qualidade e integridade dos dados***. A qualidade dos conjuntos de dados utilizados é fundamental para o desempenho dos sistemas de IA. Quando são recolhidos, os dados podem conter enviesamentos socialmente construídos, inexatidões, erros e enganos. Esta questão tem de ser resolvida antes de se treinar o sistema com um determinado conjunto de dados. Além disso, há que assegurar a integridade dos dados. A introdução de dados maliciosos num sistema de IA pode alterar o seu comportamento, em especial no caso dos sistemas com autoaprendizagem. Os processos e conjuntos de dados utilizados devem ser testados e documentados em cada uma das etapas, nomeadamente de planeamento, treino, ensaio e implantação. O mesmo se aplica aos sistemas de IA que não foram desenvolvidos a nível interno, mas sim adquiridos externamente.

74) ***Acesso aos dados***. Em qualquer organização que trate dados pessoais (independentemente de pertencerem a um utilizador ou a um não utilizador do sistema), devem ser adotados protocolos de governação do acesso aos dados. Estes protocolos devem indicar quem pode aceder aos dados e em que circunstâncias o pode fazer. O acesso a dados pessoais só deverá ser permitido a pessoal devidamente qualificado, que tenha competência e necessidade de aceder aos mesmos.

4. Transparência

75) Este requisito está estreitamente relacionado com o princípio da explicabilidade e abrange a transparência dos elementos relevantes para um sistema de IA: os dados, o sistema e os modelos de negócio.

76) ***Rastreabilidade***. Os conjuntos de dados e os processos que produzem a decisão do sistema de IA, incluindo os processos de recolha e etiquetagem dos dados, bem como os algoritmos utilizados, devem ser documentados

[40] Tratam-se de ficheiros que replicam cada etapa do processo de desenvolvimento do sistema de IA, desde a investigação e da recolha inicial de dados até aos resultados.

[41] É de referir a legislação em matéria de privacidade atualmente em vigor, como o RGPD, ou o futuro Regulamento Privacidade e Comunicações Eletrónicas.

da melhor forma possível para permitir a rastreabilidade e um aumento da transparência. Isto também se aplica às decisões tomadas pelo sistema de IA. Deste modo, é possível identificar os motivos por que uma decisão de IA foi errada, o que, por sua vez, poderá ajudar a evitar erros futuros. A rastreabilidade facilita, assim, a auditabilidade e a explicabilidade.

77) ***Explicabilidade.*** A explicabilidade diz respeito à capacidade de explicar tanto os processos técnicos de um sistema de IA como as decisões humanas com eles relacionadas (p. ex., os domínios de aplicação de um sistema de IA). A explicabilidade técnica exige que as decisões tomadas por um sistema de IA possam ser compreendidas e rastreadas por seres humanos. Além disso, poderá ser necessário adotar soluções de compromissos entre o reforço da explicabilidade de um sistema (o que poderá reduzir a sua exatidão) ou o aumento da sua exatidão (à custa da sua explicabilidade). Sempre que um sistema de IA tenha um impacto significativo na vida das pessoas, deverá ser possível solicitar uma explicação adequada do respetivo processo de tomada de decisões. Tal explicação deve ser oportuna e adaptada ao nível de especialização da parte interessada em causa (p. ex., leigo, regulador ou investigador). Além disso, devem ser disponibilizadas explicações sobre o grau de influência e de intervenção de um sistema de IA no processo decisório da organização, as opções de conceção do sistema e os fundamentos da sua implantação (assegurando assim a transparência do modelo de negócio).

78) ***Comunicação.*** Os sistemas da IA não se devem apresentar como seres humanos aos utilizadores; os seres humanos têm direito a serem informados de que estão a interagir com um sistema de IA. Tal implica que os sistemas de IA devem ser identificáveis como tal. Além disso, deve ser facultada a opção de decidir contra essa interação a favor da interação humana, sempre que necessário, a fim de garantir que os direitos fundamentais são respeitados. Além disso, as capacidades e limitações do sistema de IA devem ser comunicadas aos profissionais no domínio da IA ou aos utilizadores finais de forma adequada ao caso de utilização em questão. Essa comunicação poderá incluir o nível de exatidão do sistema de IA, bem como as suas limitações.

5. Diversidade, não discriminação e equidade

79) A inclusão e a diversidade têm de estar presentes em todo o ciclo de vida do sistema de IA para que a IA de confiança se torne uma realidade. Além da consideração e do envolvimento de todas as partes interessadas ao longo do processo, tal implica também que a igualdade de acesso mediante processos de conceção inclusivos e a igualdade de tratamento sejam asseguradas. Este requisito está estreitamente relacionado com o princípio da equidade.

80) ***Prevenção de enviesamentos injustos***. Os conjuntos de dados utilizados pelos sistemas de IA (tanto para treino como para funcionamento) podem ser afetados pela inclusão de desvios históricos inadvertidos, bem como por lacunas e por maus modelos de governação. A manutenção de tais desvios pode dar origem a discriminação e preconceitos (in)diretos não intencionais[42] contra determinados grupos ou pessoas, agravando o preconceito e a marginalização. A exploração intencional de preconceitos já existentes (entre os consumidores) e as práticas de concorrência desleal, tais como a homogeneização dos preços através de conluios ou da falta de transparência do mercado, também podem causar danos[43]. O enviesamento identificável e discriminatório deve ser eliminado na fase de recolha de dados, sempre que possível. A forma como os sistemas de IA são desenvolvidos (p. ex., a programação de algoritmos) também pode ser afetada por um enviesamento injusto. Tal pode ser combatido mediante a adoção de processos de supervisão para analisar e abordar a finalidade, os condicionalismos, os requisitos e as decisões do sistema de forma clara e

[42] Para uma definição de discriminação direta e indireta, ver, por exemplo, o artigo 2.º da Diretiva 2000/78/CE do Conselho, de 27 de novembro de 2000, que estabelece um quadro geral de igualdade de tratamento no emprego e na atividade profissional. Ver também o artigo 21.º da Carta dos Direitos Fundamentais da UE.

[43] Cf. artigo da Agência dos Direitos Fundamentais da União Europeia: «BigData: Discrimination in data-supported decision making (2018)» http://fra.europa.eu/en/publication/2018/big-data-discrimination.http://fra.europa.eu/en/publication/2018/big-data-discrimination.

transparente. Além disso, o recrutamento de pessoal de diferentes origens, culturas e disciplinas pode assegurar a diversidade de opiniões e deve ser incentivado.

81) ***Acessibilidade e conceção universal***. Nos domínios das relações entre as empresas e os consumidores, em especial, os sistemas devem centrar-se no utilizador e ser concebidos por forma a permitir que todas as pessoas utilizem os produtos ou serviços de IA, independentemente da sua idade, do seu género, das suas capacidades ou das suas características. A possibilidade de acesso a esta tecnologia por parte das pessoas com deficiência, presentes em todos os grupos da sociedade, reveste-se de especial importância. Os sistemas de IA não devem seguir uma abordagem única para todos os casos e devem tomar em consideração os princípios de conceção universal[44], que visam abranger a maior variedade possível de utilizadores, respeitando as normas de acessibilidade pertinentes[45]. Possibilitar-se-á, assim, um acesso equitativo e uma participação ativa de todas as pessoas em atividades humanas já existentes ou emergentes que utilizam computadores, nomeadamente em tecnologias de apoio[46].

82) ***Participação das partes interessadas***. Para desenvolver sistemas de IA dignos de confiança, é aconselhável consultar as partes interessadas que podem ser afetadas de forma direta ou indireta pelo sistema ao longo do seu ciclo de vida. É conveniente solicitar uma transmissão regular de observações, mesmo após a implantação, e criar mecanismos a mais longo prazo para a participação das partes interessadas, por exemplo assegurando a informação, a consulta e a participação dos trabalhadores ao longo de todo o processo de adoção dos sistemas de IA nas organizações.

6. Bem-estar societal e ambiental

83) Em conformidade com os *princípios da equidade* e *da prevenção de danos*, a sociedade em geral, outros seres sensíveis e o ambiente também devem ser considerados partes interessadas ao longo do ciclo de vida da IA. A sustentabilidade e a responsabilidade ecológica dos sistemas de IA devem ser incentivadas e deve ser promovida a investigação em soluções de IA direcionadas para áreas de interesse global, como, por exemplo, os Objetivos de Desenvolvimento Sustentável. Idealmente, a IA deve ser utilizada em benefício de todos os seres humanos, incluindo as gerações futuras.

84) ***IA sustentável e respeitadora do ambiente***. Os sistemas de IA prometem ajudar a dar resposta a algumas das preocupações societais mais prementes, mas deve assegurar-se que essa resposta é dada da forma mais respeitadora do ambiente possível. O processo de desenvolvimento, implantação e utilização do sistema, bem como toda a sua cadeia de abastecimento, deve ser avaliado a este respeito, nomeadamente através de um exame crítico da utilização de recursos e do consumo de energia durante o treino, optando-se por escolhas menos prejudiciais. As medidas destinadas a assegurar que toda a cadeia de abastecimento do sistema de IA respeita o ambiente devem ser incentivadas.

85) ***Impacto social***. A exposição omnipresente a sistemas sociais de IA[47] em todas as áreas da nossa vida (seja na educação, no trabalho, nos cuidados ou no entretenimento) pode alterar a nossa conceção de ação social ou afetar as nossas relações e laços sociais. Embora os sistemas de IA possam ser utilizados para reforçar as competências sociais[48], também podem contribuir para a sua deterioração, o que também pode afetar o

---

[44] O artigo 42.º da Diretiva Contratos Públicos exige que as especificações técnicas tomem em consideração a acessibilidade e a conceção para todos os utilizadores.

[45] Por exemplo, a norma EN 301 549.

[46] Este requisito está relacionado com a Convenção das Nações Unidas sobre os Direitos das Pessoas com Deficiência.

[47] Tal refere-se à comunicação e interação de sistemas de IA com seres humanos por via da simulação da sociabilidade na interação entre seres humanos e robôs (IA corporizada) ou como avatares na realidade virtual. Ao procederem desse modo, esses sistemas são suscetíveis de mudar as nossas práticas socioculturais e a estrutura da nossa vida social.

[48] Ver, por exemplo, o projeto financiado pela UE que desenvolve software baseado em IA que permite a robôs interagirem de forma mais eficaz com crianças autistas em sessões terapêuticas conduzidas por seres humanos, ajudando a melhorar as suas competências sociais e de comunicação:

bem-estar físico e mental das pessoas. Por conseguinte, os efeitos destes sistemas devem ser cuidadosamente acompanhados e tomados em consideração.

86) ***Sociedade e democracia.*** Além de se avaliar o impacto do desenvolvimento, da implantação e da utilização de um sistema de IA nos indivíduos, também se deverá avaliar esse impacto numa perspetiva societal, tendo em conta o seu efeito nas instituições, na democracia e na sociedade em geral. A utilização de sistemas de IA deve ser cuidadosamente ponderada, em especial em situações relacionadas com o processo democrático, incluindo não só o processo de tomada de decisões políticas, mas também os contextos eleitorais.

7. Responsabilização

87) O requisito de responsabilização complementa os requisitos acima enunciados, estando estreitamente relacionado com o *princípio da equidade*. Exige que sejam criados mecanismos para garantir a responsabilidade e a responsabilização pelos sistemas de IA e os seus resultados, tanto antes como depois da sua adoção.

88) ***Auditabilidade.*** A auditabilidade implica que seja possibilitada a avaliação de algoritmos, dados e processos de conceção. Tal não implica necessariamente que as informações sobre os modelos de negócios e a propriedade intelectual relacionadas com o sistema de IA tenham de estar sempre publicamente disponíveis. A avaliação por auditores internos e externos e a disponibilidade desses relatórios de avaliação podem contribuir para a fiabilidade da tecnologia. Em aplicações que afetem os direitos fundamentais, incluindo aplicações críticas para a segurança, os sistemas de IA devem poder ser objeto de auditorias independentes.

89) ***Minimização e comunicação dos impactos negativos***. Deve ser assegurada a capacidade de comunicar as ações ou decisões que contribuem para um determinado resultado do sistema, bem como de responder às consequências desse resultado. A identificação, a avaliação, a comunicação e a minimização dos potenciais impactos negativos dos sistemas de IA são particularmente cruciais para as pessoas (in)diretamente afetadas. Deve disponibilizar-se a devida proteção aos denunciantes, às ONG, aos sindicatos ou a outras entidades que denunciem preocupações legítimas com um sistema baseado na IA. O recurso a avaliações de impacto (p. ex., *red teaming* — simulação de ataques — ou formas de avaliação do impacto algorítmico) tanto antes como durante o desenvolvimento, a implantação e a utilização de sistemas de IA pode ser útil para minimizar os impactos negativos. Estas avaliações devem ser proporcionadas em relação ao risco colocado pelos sistemas de IA.

90) ***Soluções de compromisso.*** É possível que surjam conflitos entre os requisitos acima referidos, durante a sua aplicação, o que poderá levar à inevitabilidade da adoção de soluções de compromisso. Essas soluções de compromisso devem ser abordadas de forma racional e metodológica de acordo com os conhecimentos atuais. Tal implica que os interesses e valores pertinentes envolvidos no sistema de IA sejam identificados e que, se surgirem conflitos, as soluções de compromisso entre eles sejam explicitamente reconhecidas e avaliadas em termos do seu risco para os princípios éticos, incluindo os direitos fundamentais. Nas situações em que não seja possível identificar quaisquer soluções de compromisso eticamente aceitáveis, o desenvolvimento, a implantação e a utilização do sistema de IA não devem prosseguir dessa forma. Qualquer decisão sobre a solução de compromisso a adotar deverá ser bem fundamentada e adequadamente documentada. O decisor político tem de ser responsabilizado pela forma como a solução de compromisso adequada é formulada e deve rever continuamente a adequação da decisão resultante, para assegurar que podem introduzir-se as alterações necessárias no sistema sempre que apropriado. [49]

---

http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

[49] Há diferentes modelos de governação que podem contribuir para o efeito. Por exemplo, a presença de um perito ou painel interno e/ou externo em questões éticas (e setoriais) poderá ser útil para destacar as áreas de potencial conflito e sugerir as melhores formas de o

91) ***Vias de recurso***. Quando ocorrer um impacto adverso injusto, deverão ser previstos mecanismos acessíveis para assegurar vias de recurso adequadas[50]. Saber que é possível obter uma via de recurso quando as coisas correm mal é fundamental para garantir a confiança. Deve prestar-se especial atenção a pessoas ou grupos vulneráveis.

2. Métodos técnicos e não técnicos para concretizar uma IA de confiança

92) Podem utilizar-se métodos técnicos e não técnicos para aplicar os requisitos acima referidos. Estes métodos abrangem todas as fases do ciclo de vida de um sistema de IA. Deve realizar-se uma avaliação contínua dos métodos utilizados para aplicar os requisitos, bem como para comunicar e justificar[51] as alterações feitas aos processos de aplicação. Dado que os sistemas de IA evoluem continuamente e atuam num ambiente dinâmico, a concretização de uma IA de confiança é um processo contínuo, representado na figura 3 abaixo.

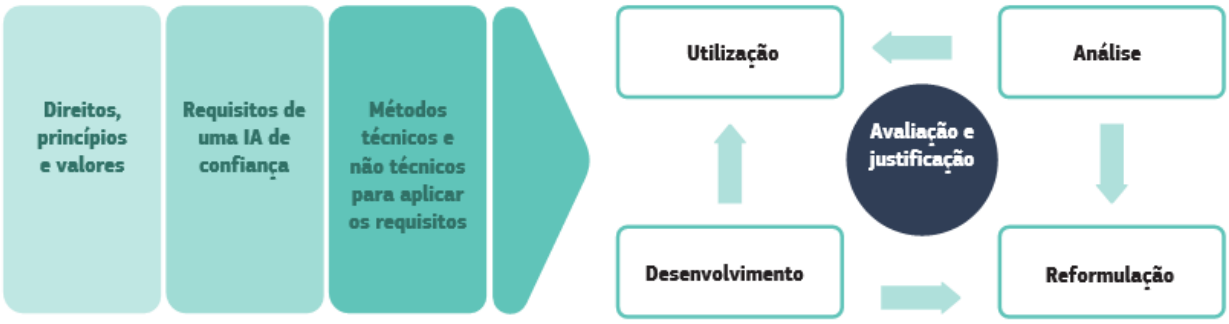

Figura 3: Concretizar uma IA de confiança ao longo de todo o ciclo de vida do sistema

93) Os métodos seguintes podem ser considerados complementares ou alternativos entre si, uma vez que diferentes requisitos — e diferentes sensibilidades — podem suscitar a necessidade de métodos de aplicação diferentes. A presente síntese não pretende ser abrangente, exaustiva ou obrigatória. O seu objetivo é, pelo contrário, oferecer uma lista de métodos sugeridos que possam ajudar a concretizar uma IA de confiança.

1. Métodos técnicos

94) Nesta secção, descrevem-se métodos técnicos para assegurar uma IA de confiança que podem ser incorporados nas fases de conceção, desenvolvimento e utilização de um sistema de IA. Os métodos abaixo enumerados variam quanto ao nível de maturidade[52].

- *Arquiteturas para uma IA de confiança*

95) Os requisitos para uma IA de confiança devem ser «traduzidos» em procedimentos e/ou restrições aos

resolver. Também é útil proceder a uma consulta e um debate significativos com as partes interessadas, incluindo aquelas que correm o risco de serem negativamente afetadas por um sistema de IA. As universidades europeias devem assumir um papel de liderança na formação dos peritos em ética necessários.

[50] Ver também o parecer da Agência dos Direitos Fundamentais da União Europeia intitulado «Improving access to remedy in the area of business and human rights at the EU level» (2017), https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] Este envolve, por exemplo, uma justificação das escolhas efetuadas na conceção, no desenvolvimento e na implantação do sistema, a fim de incorporar os requisitos acima referidos.

[52] Embora alguns deles já estejam disponíveis, outros há que ainda exigem uma investigação mais aprofundada. Os domínios em que ainda é necessário aprofundar a investigação servirão também para fundamentar o segundo documento a produzir pelo GPAN IA, ou seja, as recomendações políticas e de investimento.

procedimentos, incorporados na arquitetura do sistema de IA. Para o efeito poder-se-á adotar uma «lista branca» com um conjunto de regras (comportamentos ou estados) que o sistema deve sempre seguir, uma «lista negra» de restrições a comportamentos ou estados que o sistema nunca deve transgredir, e combinações das mesmas ou outras garantias demonstráveis mais complexas respeitantes ao comportamento do sistema. O controlo do cumprimento destas restrições durante as operações pode ser efetuado por meio de um processo separado.

96) Os sistemas de IA com capacidades de aprendizagem, que conseguem adaptar dinamicamente o seu comportamento, podem ser entendidos como sistemas não determinísticos suscetíveis de apresentar um comportamento inesperado. Estes são frequentemente analisados segundo a perspetiva teórica de um ciclo de «perceção-planeamento-ação». A adaptação desta arquitetura para garantir uma IA de confiança exige que os requisitos sejam integrados nas três etapas do ciclo: i) na etapa de «perceção», o sistema deve ser desenvolvido de modo a reconhecer todos os elementos ambientais necessários para assegurar que os requisitos são respeitados, ii) na etapa de «planeamento», o sistema apenas deve ponderar planos que cumpram os requisitos, iii) na etapa de «ação», as ações do sistema devem restringir-se aos comportamentos que cumprem os requisitos.

97) A arquitetura acima delineada é genérica e oferece apenas uma descrição imperfeita da maioria dos sistemas de IA. No entanto, apresenta pontos de ancoragem para restrições e políticas que devem refletir-se em módulos específicos para dar lugar a um sistema global digno de confiança e percecionado como tal.

- *Ética e Estado de direito desde a conceção (*X-by-design*)*

98) Métodos para garantir que os valores desde a conceção oferecem ligações precisas e explícitas entre os princípios abstratos que o sistema é obrigado a cumprir e as decisões de aplicação específicas. A ideia de que o cumprimento das normas pode ser incorporado na conceção do sistema de IA é fundamental para estes métodos. As empresas são responsáveis por identificar o impacto dos seus sistemas de IA desde o início, bem como as normas que esses sistemas devem cumprir para evitar impactos negativos. Diferentes conceitos «desde a conceção» são já amplamente utilizados, como por exemplo os de *privacidade desde a conceção* e *segurança desde a conceção*. Tal como referido acima, para conquistar a confiança, a IA necessita de ter processos, dados e resultados seguros, e deve ser concebida de modo a resistir solidamente a dados e ataques antagónicos. Deverá incluir um mecanismo de paragem à prova de falha e permitir que o funcionamento seja retomado após uma paragem forçada (p. ex., um ataque).

- *Métodos de explicação*

99) Para um sistema ser digno de confiança, temos de ser capazes de compreender por que razão se comportou de determinada forma e produziu determinada interpretação. A IA explicável (Explainable AI — XAI) é um domínio de investigação totalmente dedicado a esta questão, visando obter uma melhor compreensão dos mecanismos subjacentes ao sistema e encontrar soluções. Este é, atualmente, um desafio em aberto no caso dos sistemas de IA baseados em redes neuronais. Os processos de treino com redes neuronais podem dar origem a parâmetros de rede com valores numéricos difíceis de correlacionar com os resultados. Além disso, por vezes, pequenas alterações nos valores dos dados podem causar alterações drásticas na sua interpretação, levando um sistema, por exemplo, a confundir um autocarro escolar com uma avestruz. Esta vulnerabilidade também pode ser explorada nos ataques ao sistema. Os métodos que envolvem a investigação XAI são essenciais não só para explicar o comportamento do sistema aos seus utilizadores, mas também para implantar uma tecnologia fiável.

- *Testes e validação*

100) Devido à natureza não determinística e dependente dos contextos dos sistemas de IA, os testes tradicionais não são suficientes. As falhas dos conceitos e representações utilizados pelo sistema podem manifestar-se apenas quando um programa é aplicado a dados suficientemente realistas. Por conseguinte, para verificar e validar o tratamento dos dados, a estabilidade, a solidez e o funcionamento do modelo subjacente devem ser

cuidadosamente monitorizados, dentro de limites bem compreendidos e previsíveis, tanto durante a fase de treino como durante a implantação. Tem de ser garantido que o resultado do processo de planeamento é coerente com os dados de entrada e que as decisões são tomadas de modo a permitir a validação do processo subjacente.

101) Os testes e a validação do sistema devem ser realizados o mais cedo possível, garantindo que o sistema se comporta da forma prevista ao longo de todo o seu ciclo de vida e, em especial, após a implantação. Devem incluir todas as componentes de um sistema de IA, incluindo os dados, os modelos pré-treinados, os ambientes e o comportamento do sistema em geral, e devem ser concebidos e executados por um grupo de pessoas o mais diversificado possível. Devem desenvolver-se múltiplos critérios para analisar as categorias testadas segundo diferentes perspetivas. Poderá ponderar-se a realização de testes antagónicos por «*red teams*» fiáveis e diversificadas, que tentem deliberadamente «penetrar» no sistema para encontrar vulnerabilidades, e a oferta de «*bug bounties*» que incentivam pessoas estranhas ao sistema a detetarem e comunicarem de forma responsável os erros e fragilidades do mesmo. Por último, deve assegurar-se que os seus resultados ou ações são coerentes com os resultados dos processos precedentes, comparando-os com as políticas previamente definidas para garantir que não são violadas.

- *Indicadores de qualidade de serviço*

102) Podem definir-se indicadores adequados de qualidade de serviço para os sistemas de IA a fim de assegurar que existe um entendimento de base sobre se estes foram testados e desenvolvidos à luz de considerações de segurança e proteção. Estes indicadores podem incluir medidas para avaliar os testes e o treino dos algoritmos, bem como os parâmetros tradicionais de avaliação de *software*: funcionalidade; desempenho; usabilidade; fiabilidade; segurança; manutenibilidade.

2. Métodos não técnicos

103) Esta secção descreve uma variedade de métodos não técnicos que podem ter um papel importante para assegurar e manter uma IA de confiança. Também estes devem ser avaliados **de forma contínua**.

- *Regulamentação*

104) Como referido, já existe regulamentação para apoiar a fiabilidade da IA — veja-se a legislação em matéria de segurança dos produtos e os quadros em matéria de responsabilidade. Visto considerarmos que a regulamentação pode necessitar de ser revista, adaptada ou introduzida, simultaneamente enquanto salvaguarda e facilitadora, esta questão será retomada no segundo documento a elaborar, relativo a recomendações políticas e de investimento no domínio da IA.

- *Códigos de conduta*

105) As organizações e as partes interessadas podem subscrever as orientações e adaptar a sua carta de responsabilidade social, indicadores essenciais de desempenho, os seus códigos de conduta ou documentos de política interna para contribuírem para os esforços no sentido de uma IA de confiança. Uma organização que trabalhe num sistema de IA pode, de um modo mais geral, documentar as suas intenções, bem como fortalecê-las com normas relativas a certos valores desejáveis, como os direitos fundamentais, a transparência e a prevenção de danos.

- *Normalização*

106) As normas aplicáveis, por exemplo, à conceção, ao fabrico e às práticas empresariais podem funcionar como um sistema de gestão da qualidade para os utilizadores de IA, os consumidores, as organizações, as instituições de investigação e os governos ao oferecerem a capacidade de reconhecer e incentivar uma conduta ética nas suas decisões de compra. Além das normas convencionais, existem abordagens de corregulação: sistemas de acreditação, códigos deontológicos das profissões ou normas de conceção

conformes com os direitos fundamentais. Exemplos atuais são, designadamente, as normas ISO ou a série de normas IEEE P7000, mas no futuro poderá ser adequado adotar um eventual rótulo de «IA de confiança», o qual confirme, por referência a normas técnicas específicas, que o sistema respeita, por exemplo, os critérios de segurança, solidez técnica e explicabilidade.

- *Certificação*

107) Dado não ser expectável que todos consigam compreender plenamente o funcionamento e os efeitos dos sistemas de IA, há que prestar mais atenção às organizações que possam atestar perante o público em geral que um sistema de IA é transparente, responsável e equitativo[53]. Estas certificações aplicarão normas concebidas para diferentes domínios de aplicação e técnicas de IA, adequadamente harmonizadas com as normas setoriais e societais dos diferentes contextos. Todavia, a certificação nunca poderá substituir a responsabilidade. Por conseguinte, deverá ser complementada por quadros de responsabilização, incluindo declarações de exoneração de responsabilidade, bem como mecanismos de revisão e reparação[54].

- *Responsabilização por meio de quadros de governação*

108) As organizações devem criar quadros de governação, tanto internos como externos, que garantam a responsabilização pelas dimensões éticas das decisões associadas ao desenvolvimento, à implantação e à utilização da IA. Tal poderá incluir, por exemplo, a nomeação de uma pessoa responsável pelas questões éticas relativas à IA, ou um painel ou conselho ético interno ou externo. Entre as possíveis funções dessa pessoa, painel ou conselho, figura a supervisão e o aconselhamento. Como foi acima referido, as especificações e/ou organismos de certificação também podem ter um papel a desempenhar para este fim. Devem ser assegurados canais de comunicação com a indústria e/ou com os grupos de supervisão política, para partilhar as melhores práticas, debater dilemas ou comunicar questões emergentes que suscitam preocupações éticas. Esses mecanismos podem complementar, mas não substituir, a supervisão jurídica (p. ex., sob a forma da nomeação de um responsável pela proteção de dados ou de medidas equivalentes, exigidas por força da legislação em matéria de proteção de dados).

- *Educação e sensibilização para promover uma mentalidade ética*

109) A IA de confiança encoraja a participação esclarecida de todas as partes interessadas. A comunicação, a educação e a formação desempenham um papel importante tanto para assegurar uma ampla difusão dos conhecimentos sobre o potencial impacto dos sistemas de IA como para sensibilizar as pessoas para o facto de que podem influenciar o desenvolvimento societal. Incluem-se aqui todas as partes interessadas, por exemplo, as pessoas envolvidas no fabrico de produtos (responsáveis pela conceção e desenvolvimento), os utilizadores (empresas ou indivíduos) e outros grupos afetados (aqueles que não podem comprar ou utilizar um sistema de IA, mas que são visados pelas decisões tomadas por um sistema de IA, e a sociedade em geral). Deve promover-se uma literacia básica no domínio da IA em toda a sociedade. Um pré-requisito para educar o público é assegurar que existem nesse espaço especialistas em ética com as competências e a formação adequadas.

- *Participação das partes interessadas e diálogo social*

110) Os benefícios da IA são muitos e a Europa necessita de assegurar que estão à disposição de todos. Tal exige um debate aberto e o envolvimento dos parceiros sociais, das partes interessadas e do público em geral. Muitas organizações já recorrem a painéis constituídos por partes interessadas para debater a utilização de sistemas de IA e análise de dados. Estes painéis são constituídos por vários membros, designadamente especialistas em

[53] Tal como defende, por exemplo, a Iniciativa de Conceção Ética (Ethically Aligned Design Initiative) da IEEE: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Para saber mais sobre as limitações da certificação, consultar: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

questões jurídicas, técnicas e éticas, representantes dos consumidores e dos trabalhadores. Procurar ativamente a participação e o diálogo sobre a utilização e o impacto dos sistemas de IA contribui para a avaliação dos resultados e das abordagens e pode ser particularmente útil nos casos mais complexos.

- *Diversidade e equipas de conceção inclusiva*

111) A diversidade e a inclusão desempenham um papel essencial no desenvolvimento dos sistemas de IA que serão utilizados no mundo real. É essencial que, à medida que os sistemas de IA executam mais tarefas de forma autónoma, as equipas que concebem, desenvolvem, testam, mantêm, implantam e/ou compram estes sistemas reflitam a diversidade dos utilizadores e da sociedade em geral. Tal contribui para a objetividade e para a consideração de diferentes perspetivas, necessidades e objetivos. O ideal é que as equipas sejam diversificadas não só em termos de género, cultura e idade, mas também em termos de experiências profissionais e conjuntos de competências.

Orientações fundamentais extraídas do capítulo II:

- ✓ Assegurar que todo o ciclo de vida do sistema de IA cumpre os requisitos para uma IA de confiança: 1) ação e supervisão humanas; 2) solidez técnica e segurança; 3) privacidade e governação dos dados; 4) transparência; 5) diversidade, não discriminação e equidade; 6) bem-estar ambiental e societal; 7) responsabilização.
- ✓ Ponderar métodos técnicos e não técnicos para assegurar a aplicação desses requisitos.
- ✓ Promover a investigação e a inovação para ajudar a avaliar os sistemas de IA e a melhorar o cumprimento dos requisitos; divulgar os resultados e as questões em aberto junto do público em geral e formar sistematicamente uma nova geração de peritos em ética associada à IA.
- ✓ Comunicar, de forma clara e proativa, informações às partes interessadas sobre as capacidades e as limitações do sistema de IA, permitindo-lhes criar expectativas realistas, e sobre a forma como os requisitos são aplicados. Ser transparente sobre o facto de estarem a lidar com um sistema de IA.
- ✓ Facilitar a rastreabilidade e a auditabilidade dos sistemas de IA, sobretudo em contextos ou situações críticos.
- ✓ Envolver as partes interessadas em todo o ciclo de vida do sistema de IA. Promover a formação e a educação para que todas as partes interessadas tenham conhecimento e recebam formação em matéria de IA de confiança.
- ✓ Estar ciente de que podem existir conflitos fundamentais entre diferentes princípios e requisitos. Identificar, avaliar, documentar e comunicar continuamente essas soluções de compromisso.

III. Capítulo III: Avaliação de uma IA de confiança

112) Com base nos requisitos essenciais expressos no capítulo II, este capítulo estabelece uma **lista de avaliação de uma IA de confiança** não exaustiva (versão piloto) para **operacionalizar uma IA de confiança**. Esta lista é aplicável, em especial, a sistemas de IA que interagem diretamente com os utilizadores e destina-se principalmente aos criadores e implantadores de sistemas de IA (desenvolvidos por eles próprios ou adquiridos a terceiros). Esta lista de avaliação não aborda a operacionalização da primeira componente de uma IA de confiança (IA legal). A conformidade com a presente lista de avaliação não constitui prova de conformidade legal, nem a lista se destina a servir de orientação para garantir o cumprimento da legislação aplicável. Tendo em conta a especificidade da aplicação dos sistemas de IA, a lista de avaliação terá de ser adaptada aos casos de utilização e contextos específicos em que os sistemas funcionam. Além disso, o presente capítulo oferece uma recomendação geral sobre a forma de adotar a lista de avaliação de uma IA de confiança por meio de uma estrutura de governação que abarca tanto o nível operacional como o nível de gestão.

113) A lista de avaliação e a estrutura de governação serão desenvolvidas em estreita colaboração com as partes interessadas dos setores público e privado. Este desenvolvimento será conduzido como um processo-piloto que permita a obtenção alargada de observações mediante dois processos paralelos:

a. Um processo qualitativo que assegure a representatividade, mediante o qual uma pequena seleção de empresas, organizações e instituições (de diferentes setores e dimensões) se inscreverão para testar a lista de avaliação e a estrutura de governação na prática, bem como para fornecer observações circunstanciadas sobre essa experiência;

b. Um processo quantitativo em que todas as partes interessadas se podem inscrever para testar a lista de avaliação e apresentar observações sobre essa experiência no âmbito de uma consulta pública.

114) Após a fase-piloto, integraremos os resultados do processo de obtenção de observações na lista de avaliação e elaboraremos uma versão revista da mesma no início de 2020. O objetivo é obter um quadro que possa ser transversalmente utilizado em todas as aplicações e oferecer, assim, uma base para assegurar uma IA de confiança em todos os domínios. Depois de se estabelecer essa base, será possível desenvolver um quadro setorial ou específico de cada aplicação.

Governação

115) As empresas, organizações e instituições poderão querer analisar as possíveis formas de aplicar a lista de avaliação de uma IA de confiança. Para o efeito poder-se-á incluir o processo de avaliação nos mecanismos de governação existentes ou adotar novos processos. Tal escolha dependerá da estrutura interna da organização, bem como da sua dimensão e dos recursos disponíveis.

116) A investigação[55] demonstra que é essencial uma atenção dos órgãos de gestão ao mais alto nível para que a mudança seja possível. Demonstra igualmente que o envolvimento de todos os interessados numa empresa, organização ou instituição fomenta a aceitação e a pertinência da introdução de qualquer processo novo (independentemente de ser tecnológico ou não)[56]. Por conseguinte, recomendamos a aplicação de um processo que procure envolver tanto o nível operacional como o nível de gestão de topo.

Nível	**Funções pertinentes (dependendo da organização)**
Conselho de Administração	A gestão de topo debate e avalia o desenvolvimento, a implantação ou a aquisição de IA e constitui a instância superior para avaliar todas as inovações e utilizações de IA, quando são detetadas preocupações críticas. Envolve as pessoas afetadas pela eventual introdução de sistemas de IA (p. ex., os trabalhadores) e os seus representantes ao longo de todo o processo, através de procedimentos de informação, consulta e participação.
Departamento jurídico/de conformidade/de responsabilidade social	O departamento de responsabilidade controla a utilização da lista de avaliação e a sua necessária evolução para acompanhar as mudanças tecnológicas e regulamentares. Atualiza as normas ou políticas internas relativas aos sistemas de IA e garante que a utilização de tais sistemas está conforme com o quadro legal e regulamentar em vigor e com os valores da organização.

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation.

[56] Ver, por exemplo, A. Bryson, E. Barth and H. Dale-Olsen, «The Effects of Organisational change on worker well-being and the moderating role of trade unions», ILRReview, 66(4), julho de 2013; Jirjahn, U. e Smith, S.C. (2006). «What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations», 45(4), 650–680; Michie, J. e Sheehan, M. (2003). «Labour market deregulation, "flexibility" and innovation», Cambridge Journal of Economics, 27(1), 123–143.

Desenvolvimento de produtos e serviços, ou equivalente	O departamento de desenvolvimento de produtos e serviços utiliza a lista de avaliação para avaliar os produtos e serviços baseados em IA e regista todos os resultados. Estes resultados são debatidos a nível da gestão, que aprova em última instância as aplicações baseadas em IA, novas ou revistas.
Garantia da qualidade	O departamento de garantia da qualidade (ou equivalente) garante e verifica os resultados da lista de avaliação e toma medidas para levar uma questão ao nível de gestão superior, se o resultado não for satisfatório ou se forem detetados resultados imprevistos.
Recursos humanos	O departamento de recursos humanos assegura a combinação apropriada de competências e a diversidade de perfis dos criadores de sistemas de IA. Assegura que é ministrado o nível de formação adequado sobre a IA de confiança dentro da organização.
Contratação pública	O departamento de contratação pública assegura que o processo de aquisição de produtos ou serviços baseados em IA inclui uma verificação da IA de confiança.
Operações correntes	Os criadores e gestores de projetos incluem a lista de avaliação no seu trabalho quotidiano e documentam os resultados e as conclusões da avaliação.

Utilização da lista de avaliação de uma IA de confiança

117) Ao utilizar a lista de avaliação na prática, recomendamos que se preste atenção não só aos domínios que suscitam preocupação, mas também às perguntas que não podem ser (facilmente) respondidas. Um potencial problema poderá ser a falta de diversidade das capacidades e competências da equipa que está a desenvolver e a testar o sistema de IA, podendo ser necessário envolver outras partes interessadas internas ou externas à organização. Recomenda-se vivamente o registo de todos os resultados, tanto em termos técnicos como em termos de gestão, assegurando que a resolução de problemas pode ser compreendida a todos os níveis da estrutura de governação.

118) A presente lista de avaliação destina-se a orientar todos os profissionais no domínio da IA no desenvolvimento, na implantação e na utilização de uma IA de confiança. A avaliação deve ser adaptada ao caso de utilização específica de uma forma proporcionada. Durante a fase-piloto, poderão ser reveladas áreas sensíveis, sendo a necessidade de mais especificações em tais casos avaliada na etapa seguinte. Embora esta lista de avaliação não dê respostas concretas para as questões suscitadas, incentiva a reflexão sobre as medidas que podem ajudar a garantir a fiabilidade dos sistemas de IA e os possíveis passos a dar nesse sentido.

Relação com a legislação e os processos em vigor

119) É igualmente importante que os envolvidos no desenvolvimento, na implantação e na utilização da IA reconheçam que há várias leis em vigor que obrigam à utilização de determinados processos e proíbem determinados resultados, as quais podem sobrepor-se e coincidir com algumas das medidas enumeradas na lista de avaliação. Por exemplo, a legislação em matéria de proteção de dados define uma série de requisitos legais que devem ser cumpridos pelos envolvidos na recolha e no tratamento de dados pessoais. Todavia, como uma IA de confiança também exige o tratamento ético dos dados, os procedimentos e as políticas a nível interno destinados a assegurar o cumprimento da legislação em matéria de proteção de dados também podem ajudar a facilitar a gestão ética dos dados e complementar,

assim, os processos jurídicos existentes. Todavia, a conformidade com a presente lista de avaliação *não* constitui prova de conformidade legal, nem a lista se destina a servir de orientação para garantir o cumprimento da legislação aplicável. Em vez disso, o seu intuito é apresentar um conjunto de questões específicas aos destinatários com o intuito de assegurar que a sua abordagem ao desenvolvimento e à implantação da IA é orientada no sentido de uma IA de confiança que procura concretizar.

120) Do mesmo modo, muitos profissionais no domínio da IA já têm instrumentos de avaliação e processos de desenvolvimento de software em vigor para garantir a conformidade também com normas não jurídicas. A avaliação a seguir apresentada não deverá ser necessariamente realizada como um exercício autónomo, mas pode ser incorporada nessas práticas existentes.

LISTA DE AVALIAÇÃO DE UMA IA DE CONFIANÇA (VERSÃO-PILOTO)

1. Ação e supervisão humanas

Direitos fundamentais:

- ✓ Nos casos de utilização em que poderá haver um impacto negativo nos direitos fundamentais, realizou uma avaliação de impacto nos direitos fundamentais? Identificou e documentou potenciais soluções de compromisso estabelecidas entre os diferentes princípios e direitos?
- ✓ O sistema de IA interage com a tomada de decisões por utilizadores finais humanos (p. ex., recomendação de ações ou decisões a tomar, apresentação de opções)?
  - Nesses casos, existe algum risco de que o sistema de IA afete a autonomia humana interferindo com o processo decisório do utilizador final de uma forma não intencional?
  - Ponderou se o sistema de IA deveria comunicar aos utilizadores que uma decisão, um conteúdo, um conselho ou um resultado provém de uma decisão algorítmica?
  - Caso o sistema de IA inclua um sistema de conversação automática (*chat bot*), os utilizadores finais humanos foram informados do facto de estarem a interagir com um agente não humano?

Ação humana:

- ✓ Caso o sistema de IA seja introduzido num processo de trabalho, ponderou a distribuição de tarefas entre o sistema de IA e os trabalhadores humanos no que diz respeito a interações significativas e a uma supervisão e um controlo adequados por seres humanos?
  - O sistema de IA melhora ou aumenta as capacidades humanas?
  - Adotou salvaguardas para evitar o excesso de confiança ou o excesso de dependência face ao sistema de IA nos processos de trabalho?

Supervisão humana:

- ✓ Ponderou qual seria o nível adequado de controlo humano para o sistema de IA e o caso de utilização específicos?
  - Pode descrever o nível de controlo ou envolvimento humano, se aplicável? Quem é o «ser humano no controlo» e quais são os momentos ou as ferramentas para a intervenção humana?

- Criou mecanismos e medidas para assegurar esse potencial controlo ou supervisão por seres humanos, ou para garantir que as decisões são tomadas sob a responsabilidade global de seres humanos?
- Tomou algumas medidas para permitir uma auditoria e corrigir questões relacionadas com a governação da autonomia da IA?

✓ Caso exista um sistema de IA ou caso de utilização com autoaprendizagem ou autónomo, adotou mecanismos mais específicos de controlo e de supervisão?

- Que tipo de mecanismos de deteção e de resposta estabeleceu para avaliar se algo poderia correr mal?
- Assegurou a existência de um «botão de paragem» ou um procedimento para abortar uma operação de forma segura, se necessário? Esse procedimento aborta o processo por completo, parcialmente ou delega o controlo num ser humano?

2. <u>Solidez técnica e segurança</u>

Resiliência perante ataques e segurança:

✓ Avaliou potenciais formas de ataque a que o sistema de IA poderá ser vulnerável?

- Em particular, tomou em consideração diferentes tipos e naturezas de vulnerabilidades, como a poluição de dados, as infraestruturas físicas ou os ciberataques?

✓ Adotou medidas ou sistemas para garantir a integridade e a resiliência do sistema de IA contra potenciais ataques?

✓ Avaliou como se comporta o seu sistema em situações e ambientes inesperados?

✓ Ponderou se o seu sistema podia ou não, e até que ponto, ser de dupla utilização? Em caso afirmativo, tomou medidas preventivas adequadas contra essa possibilidade (incluindo, por exemplo, a não publicação da investigação ou a não implantação do sistema)?

Plano de recurso e segurança geral:

✓ Assegurou que o seu sistema tem um plano de recurso suficiente caso se depare com ataques antagónicos ou outras situações inesperadas (p. ex., procedimentos de comutação técnica ou solicitação da intervenção de um operador humano antes de prosseguir)?

✓ Analisou o nível de risco do sistema de IA neste caso de utilização específico?

- Adotou algum processo para medir e avaliar os riscos e a segurança?
- Forneceu as informações necessárias em caso de risco para a integridade física dos seres humanos?
- Ponderou a aquisição de uma apólice de seguro para cobrir eventuais danos causados pelo sistema de IA?

- Identificou os potenciais riscos de segurança de (outras) utilizações previsíveis da tecnologia, incluindo a má utilização acidental ou malévola da mesma? Existe algum plano para atenuar ou gerir estes riscos?

✓ Avaliou se existe alguma probabilidade de o sistema de IA poder causar danos ou prejuízos aos utilizadores ou a terceiros? Em caso afirmativo, avaliou a probabilidade, os potenciais danos, o público afetado e a gravidade?

- Caso existam riscos de o sistema de IA causar danos, analisou a regulamentação em matéria de responsabilidade e de defesa do consumidor? De que modo teve essa regulamentação em conta?
- Tomou em consideração o potencial impacto ou risco para a segurança do ambiente ou de animais?
- A sua análise de risco examinou se existem problemas de segurança ou de rede (p. ex., potenciais riscos para a cibersegurança) que ponham em risco a segurança ou possam causar danos devido a um comportamento não intencional do sistema de IA?

✓ Procedeu a uma estimativa do impacto provável de uma falha do seu sistema de IA que o leve a fornecer resultados incorretos, que o torne indisponível ou que o faça fornecer resultados inaceitáveis do ponto de vista societal (p. ex., práticas discriminatórias)?

- Definiu limiares e medidas de governação, caso se verifiquem os cenários acima referidos, para acionar planos alternativos/de recurso?
- Definiu e testou planos de recurso?

Exatidão

✓ Avaliou o nível e a definição de exatidão que seriam necessários no contexto do sistema de IA e do caso de utilização?

- Avaliou o modo como a exatidão é medida e assegurada?
- Tomou medidas para assegurar que os dados utilizados são exaustivos e atualizados?
- Tomou medidas para avaliar se são necessários dados adicionais, por exemplo para melhorar a exatidão ou eliminar os enviesamentos?

✓ Avaliou os danos que podem ser causados se o sistema de IA fizer previsões incorretas?

✓ Adotou formas de medir se o seu sistema está a produzir um número inaceitável de previsões incorretas?

✓ Adotou alguma série de medidas para resolver uma situação em que estejam a ser feitas previsões incorretas?

Fiabilidade e reprodutibilidade:

✓ Adotou uma estratégia para controlar e testar se o sistema de IA cumpre os seus objetivos, finalidades e aplicações previstas?

- Testou se é necessário ter em conta contextos ou condições específicos para garantir a reprodutibilidade?

- Adotou processos ou métodos de verificação para medir e assegurar os diferentes aspetos da fiabilidade e da reprodutibilidade?
- Adotou processos para descrever quando um sistema de IA falha em alguns tipos de contextos?
- Documentou claramente e operacionalizou esses processos para testar e verificar a fiabilidade dos sistemas de IA?

  Adotou mecanismos ou formas de comunicação para garantir aos utilizadores (finais) a fiabilidade do sistema de IA?

3. <u>Privacidade e governação dos dados</u>

Respeito da privacidade e proteção dos dados:

✓ Em função do caso de utilização, criou um mecanismo que permita que outras pessoas assinalem problemas de privacidade ou proteção de dados relacionados com os processos de recolha (para treino e funcionamento) e de tratamento de dados?

✓ Avaliou o tipo e o âmbito dos dados dos seus conjuntos de dados (p. ex., se contêm dados pessoais)?

✓ Analisou formas de desenvolver o sistema de IA ou de treinar o modelo sem a utilização ou com uma utilização mínima de dados potencialmente sensíveis ou pessoais?

✓ Incorporou mecanismos para assinalar e controlar dados pessoais em função do caso de utilização (tais como o consentimento válido e a possibilidade de revogação, quando aplicável)?

✓ Tomou medidas para aumentar a privacidade, tais como a encriptação, a anonimização e a agregação?

✓ Caso exista um responsável pela proteção de dados, envolveu essa pessoa na fase inicial do processo?

Qualidade e integridade dos dados:

✓ Harmonizou o seu sistema com potenciais normas pertinentes (p. ex., ISO, IEEE) ou protocolos amplamente adotados para a sua gestão e governação quotidianas dos dados?

✓ Criou mecanismos de supervisão para a recolha, a conservação, o tratamento e a utilização de dados?

✓ Avaliou em que medida controla a qualidade das fontes de dados externas utilizadas?

✓ Adotou processos para garantir a qualidade e a integridade dos seus dados? Ponderou a adoção de outros processos? Como verifica se os seus conjuntos de dados não foram comprometidos ou objeto de pirataria informática?

Acesso aos dados:

✓ Que protocolos, processos e procedimentos foram seguidos para gerir e assegurar uma governação adequada dos dados?

- Avaliou quem pode aceder aos dados dos utilizadores e em que circunstâncias?
- Garantiu que estas pessoas são qualificadas e necessitam de aceder aos dados, e que possuem as

competências necessárias para compreender a política de proteção de dados ao pormenor?

- Assegurou um mecanismo de supervisão para registar quando, onde, como, por quem e para que fim os dados foram acedidos?

4. <u>Transparência</u>

Rastreabilidade:

✓ Adotou medidas que garantam a rastreabilidade? Tal poderá implicar a documentação de:

- Métodos utilizados para conceber e desenvolver o sistema algorítmico:
  - no caso de um sistema de IA baseado em regras, o método de programação ou a forma como o modelo foi construído devem ser documentados;
  - no caso de um sistema de IA baseado na aprendizagem, o método de treino do algoritmo, incluindo os dados de entrada que foram recolhidos e selecionados, e a forma como isso foi feito devem ser documentados.
- Métodos utilizados para testar e validar o sistema algorítmico:
  - no caso de um sistema de IA baseado em regras, os cenários ou casos utilizados para o testar e validar devem ser documentados;
  - no caso de um modelo baseado na aprendizagem, as informações sobre os dados utilizados para o testar e validar devem ser documentadas.
- Resultados do sistema algorítmico:
  - Os resultados ou as decisões tomadas pelo algoritmo, bem como outras decisões potenciais que resultariam de casos diferentes (p. ex., para outros subgrupos de utilizadores) devem ser documentados.

Explicabilidade:

✓ Avaliou em que medida as decisões e, logo, os resultados produzidos pelo sistema de IA podem ser compreendidos?

✓ Assegurou que uma explicação dos motivos por que um sistema fez determinada escolha que levou a um determinado resultado pode ser tornada compreensível para todos os utilizadores que desejem obter uma explicação?

✓ Avaliou em que medida a decisão do sistema influencia os processos de tomada de decisões da organização?

✓ Avaliou por que razão o sistema em causa foi utilizado neste domínio específico?

✓ Avaliou o modelo de negócios respeitante a este sistema (p. ex., como é que ele cria valor para a organização)?

✓ Concebeu o sistema de IA tendo a sua interpretabilidade em conta desde o início?

- Investigou e procurou utilizar o modelo mais simples e fácil de interpretar para a aplicação em questão?

- Avaliou se pode analisar os seus dados utilizados durante o treino e os testes? Pode alterá-los e atualizá-los ao longo do tempo?
- Verificou se tem algumas opções após o treino e o desenvolvimento do modelo para examinar a interpretabilidade, ou se tem acesso ao fluxo de trabalho interno do modelo?

Comunicação:

✓ Comunicou aos utilizadores (finais) — por via de uma declaração de exoneração de responsabilidade ou de qualquer outro meio — que estão a interagir com um sistema de IA e não com outro ser humano? Rotulou o seu sistema de IA como tal?

✓ Criou mecanismos para informar os utilizadores acerca das razões e dos critérios subjacentes aos resultados do sistema de IA?

- Essa informação é comunicada de forma clara e inteligível aos utilizadores previstos?
- Estabeleceu processos para ter em conta as observações dos utilizadores e utiliza-as para adaptar o sistema?
- Comunicou também os riscos potenciais ou percecionados, tais como enviesamentos?
- Em função do caso de utilização, tomou também em consideração a comunicação e a transparência face a outros públicos, a terceiros ou ao público em geral?

✓ Explicou claramente qual é a finalidade do sistema de IA e os eventuais beneficiários do produto/serviço?

- Os cenários de utilização do produto foram especificados e claramente comunicados, tomando também em consideração formas de comunicação alternativas para assegurar que o produto é compreensível e adequado para o utilizador a que se destina?
- Em função do caso de utilização, refletiu sobre a psicologia humana e as potenciais limitações, como o risco de confusão, o enviesamento da confirmação ou a fadiga cognitiva?

✓ Comunicou claramente as características, as limitações e as potenciais insuficiências do sistema de IA:

- em caso de desenvolvimento: a quem está a implantá-lo num produto ou serviço?
- em caso de implantação: ao utilizador final ou consumidor?

5. <u>Diversidade, não discriminação e equidade</u>

Prevenção de enviesamentos injustos:

✓ Assegurou uma estratégia ou um conjunto de procedimentos para evitar criar ou reforçar enviesamentos injustos no sistema de IA, no que respeita tanto à utilização de dados de entrada como à conceção do algoritmo?

- Avaliou e reconheceu as eventuais limitações decorrentes da composição dos conjuntos de dados utilizados?
- Tomou em consideração a diversidade e a representatividade dos utilizadores nos dados?

Efetuou testes em relação a populações específicas ou a casos de utilização problemáticos?

- Investigou e utilizou as ferramentas técnicas disponíveis para melhorar a sua compreensão dos dados, do modelo e do desempenho?
- Criou processos para testar e controlar potenciais enviesamentos durante as fases de desenvolvimento, implantação e utilização do sistema?

✓ Em função do caso de utilização, assegurou a existência de um mecanismo para permitir que outros assinalem questões relacionadas com enviesamento, discriminação ou mau desempenho do sistema de IA?

- Ponderou a adoção de medidas e formas de comunicação claras sobre o modo de suscitar essas questões e a quem podem ser apresentadas?
- Teve em conta não só os utilizadores (finais) mas também outras pessoas que possam ser indiretamente afetadas pelo sistema de IA?

✓ Avaliou se pode ocorrer uma eventual variabilidade das decisões em condições idênticas?

- Em caso afirmativo, analisou quais seriam as possíveis causas para tal?
- Em caso de variabilidade, estabeleceu um mecanismo de medição ou avaliação do potencial impacto dessa variabilidade nos direitos fundamentais?

✓ Estabeleceu uma definição operacional adequada de «equidade» que aplique na conceção dos sistemas de IA?

- A sua definição é comummente utilizada? Ponderou outras definições antes de escolher a que está a utilizar?
- Assegurou uma análise quantitativa ou parâmetros para medir e testar a definição de equidade utilizada?
- Estabeleceu mecanismos para garantir a equidade dos seus sistemas de IA? Ponderou utilizar outros mecanismos possíveis?

Acessibilidade e conceção universal:

✓ Assegurou que o sistema de IA abrange uma vasta gama de preferências e capacidades individuais?

- Avaliou se o sistema de IA pode ser utilizado por pessoas com necessidades especiais ou deficiência, ou pessoas em risco de exclusão? De que forma foi esta possibilidade incorporada na conceção do sistema e como é verificada?
- Assegurou que as informações sobre o sistema de IA também estão acessíveis a utilizadores de tecnologias de apoio?
- Envolveu ou consultou esta comunidade durante a fase de desenvolvimento do sistema de IA?

✓ Teve em conta o impacto do seu sistema de IA no grupo potencial de utilizadores?

- A equipa envolvida na construção do sistema de IA é representativa do seu público-alvo de utilizadores? É representativa da população em geral, considerando também outros grupos que

possam ser tangencialmente afetados?

- Avaliou se poderão existir pessoas ou grupos desproporcionadamente afetados pelas implicações negativas?
- Obteve observações de outras equipas ou outros grupos que representem diferentes contextos e experiências?

Participação das partes interessadas:

✓ Ponderou um mecanismo para incluir a participação das diferentes partes interessadas no desenvolvimento e na utilização do sistema de IA?

✓ Preparou o terreno para a introdução de um sistema de IA na sua organização informando e envolvendo previamente os trabalhadores afetados e os seus representantes?

6. <u>Bem-estar societal e ambiental</u>

IA sustentável e respeitadora do ambiente:

✓ Criou mecanismos para medir o impacto ambiental do desenvolvimento, da implantação e da utilização do sistema de IA (p. ex., energia utilizada pelo centro de dados, tipo de energia utilizada pelos centros de dados, etc.)?

✓ Assegurou a adoção de medidas para reduzir o impacto ambiental do ciclo de vida do seu sistema de IA?

Impacto social:

✓ Caso o sistema de IA interaja diretamente com seres humanos:

- Avaliou se o sistema de IA encoraja os seres humanos a desenvolver laços e empatia com o sistema?
- Assegurou que o sistema de IA assinale claramente que a sua interação social é simulada e que não tem qualquer capacidade para «compreender» ou «sentir»?

✓ Assegurou que os impactos sociais do sistema de IA são bem compreendidos? Por exemplo, avaliou se há risco de perda de postos de trabalho ou de perda de competências da mão de obra? Que medidas foram tomadas para combater tais riscos?

Sociedade e democracia:

✓ Avaliou o impacto societal mais geral da utilização do sistema de IA, para lá do utilizador (final) individual, por exemplo, sobre as partes interessadas que poderão ser indiretamente afetadas?

7. <u>Responsabilização</u>

Auditabilidade:

✓ Criou mecanismos para facilitar a auditabilidade do sistema por auditores internos e/ou independentes, designadamente para garantir a rastreabilidade e o registo dos processos e

resultados do sistema de IA?

Minimização e comunicação dos impactos negativos:

- ✓ Realizou uma avaliação de riscos ou de impacto do sistema de IA que tenha em conta as diferentes partes interessadas direta ou indiretamente afetadas?
- ✓ Criou enquadramentos de formação e educação destinados a desenvolver práticas de responsabilização?
  - Que trabalhadores ou partes da equipa estão envolvidas? Essas medidas vão além da fase de desenvolvimento?
  - Essas ações de formação também ensinam o potencial quadro jurídico aplicável ao sistema de IA?
  - Ponderou criar um «conselho de análise da IA ética» ou um mecanismo semelhante para debater as práticas deontológicas e de responsabilização em geral, incluindo zonas «cinzentas» eventualmente pouco claras?
- ✓ Complementarmente às iniciativas ou aos quadros de supervisão a nível interno das questões de ética e responsabilização, há algum tipo de orientação externa ou foram também criados processos de auditoria?
- ✓ Estão disponíveis processos para que terceiros (p. ex., fornecedores, consumidores, distribuidores/vendedores) ou trabalhadores comuniquem eventuais vulnerabilidades, riscos ou enviesamentos no sistema/aplicação de IA?

Documentação de soluções de compromisso:

- ✓ Estabeleceu um mecanismo para identificar interesses e valores pertinentes envolvidos no sistema de IA e as eventuais soluções de compromisso entre eles?
- ✓ Que processo utiliza para decidir sobre essas soluções de compromisso? Assegurou que a decisão de compromisso foi documentada?

Disponibilidade de vias de recurso:

- ✓ Estabeleceu um conjunto adequado de mecanismos de recurso na eventualidade de ocorrerem danos ou impactos negativos?
- ✓ Criou mecanismos para fornecer informações aos utilizadores (finais)/terceiros sobre as possibilidades de recurso?

Convidamos todas as partes interessadas a testarem na prática esta lista de avaliação e a apresentarem observações sobre a sua aplicabilidade, exaustividade, pertinência para a aplicação de IA ou o domínio em causa, bem como sobre a sua sobreposição ou complementaridade com os processos de conformidade ou avaliação existentes. Com base nestas observações, no início de 2020 será proposta à Comissão uma versão revista da lista de avaliação de uma IA de confiança.

Orientações fundamentais extraídas do capítulo III:

- ✓ Adotar uma **lista de avaliação** para uma IA de confiança aquando do desenvolvimento, da implantação ou da utilização da IA, e adaptá-la ao caso de utilização específico em que o sistema está a ser utilizado.
- ✓ Importa ter em mente que essa lista de avaliação **nunca será exaustiva**. Assegurar uma IA de confiança não se resume a um exercício de preenchimento de formulários; trata-se, sim, de um processo contínuo de identificação de requisitos, de avaliação de soluções e de garantia de melhores resultados ao longo do ciclo de vida do sistema de IA, e de envolvimento das partes interessadas neste processo.

C. EXEMPLOS DE OPORTUNIDADES E PREOCUPAÇÕES CRÍTICAS SUSCITADAS PELA IA

121) Na secção seguinte, fornecemos exemplos de desenvolvimento e utilização da IA que devem ser incentivados, bem como exemplos de situações em que o desenvolvimento, a implantação ou a utilização da IA pode contrariar os nossos valores e suscitar preocupações específicas. Há que encontrar o equilíbrio entre o que deve e o que pode ser feito com a IA, devendo ter-se o devido cuidado com aquilo que não deve ser feito com ela.

1. Exemplos de oportunidades oferecidas por uma IA de confiança

122) Uma IA de confiança pode constituir uma grande oportunidade para apoiar a atenuação de desafios prementes com que a sociedade está confrontada, tais como o envelhecimento da população, as crescentes desigualdades sociais e a poluição ambiental. Este potencial também está refletido a nível mundial, por exemplo nos Objetivos de Desenvolvimento Sustentável das Nações Unidas[57]. A secção seguinte examina a forma de incentivar uma estratégia europeia para a IA que dê resposta a alguns destes desafios.

a. Ação climática e infraestruturas sustentáveis

123) Embora o combate às alterações climáticas deva constituir uma prioridade máxima para os decisores políticos do mundo inteiro, a transformação digital e a IA de confiança têm grandes potencialidades para reduzir o impacto dos seres humanos no ambiente e permitir uma utilização eficiente e eficaz da energia e dos recursos naturais[58]. A IA de confiança pode ser, por exemplo, associada aos megadados para detetar as necessidades de energia com maior exatidão, possibilitando uma infraestrutura energética e um consumo de energia mais eficientes[59].

124) Em setores como os transportes públicos, os sistemas de IA para sistemas de transporte inteligentes[60] podem ser utilizados para minimizar as filas, otimizar a seleção de percursos, permitir que as pessoas com deficiência visual se tornem mais independentes[61], otimizar os motores energeticamente eficientes e reforçar, assim, os esforços de descarbonização e reduzir a pegada ambiental, em prol de uma sociedade mais ecológica. Atualmente, a nível mundial, morre uma pessoa a cada 23 segundos num acidente de viação[62]. Os sistemas de IA podem ajudar a reduzir significativamente o número mortes, por exemplo melhorando os tempos de reação

---

57 https://sustainabledevelopment.un.org/?menu=1300.

58 Vários projetos da UE têm em vista o desenvolvimento de redes inteligentes e o armazenamento de energia, que poderão contribuir para o sucesso de uma transição energética apoiada pelas tecnologias informáticas, nomeadamente através de soluções baseadas na IA e outras soluções digitais. Para complementar o trabalho desses projetos individuais, a Comissão lançou a iniciativa BRIDGE, que permite que os projetos do Programa-Quadro Horizonte 2020 em curso no domínio das redes inteligentes e do armazenamento de energia criem uma visão comum sobre questões transversais: https://www.h2020-bridge.eu/.

59 Ver, por exemplo, o projeto Encompass: http://www.encompass-project.eu/.

60 As novas soluções baseadas na IA ajudam a preparar as cidades para o futuro da mobilidade. Ver, por exemplo, o projeto financiado pela UE, denominado Fabulos: https://fabulos.eu/.

61 Ver, por exemplo, o projeto PRO4VIP, que é parte da estratégia europeia «Visão 2020» para combater a cegueira prevenível, em especial devida ao envelhecimento. A mobilidade e a orientação constituíram domínios prioritários do projeto.

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

e o cumprimento das regras[63].

b. Saúde e bem-estar

125) As tecnologias de IA de confiança podem ser utilizadas — e já o estão a ser — para tornar os tratamentos mais inteligentes e para ajudar a prevenir doenças potencialmente mortais[64]. Os médicos e outros profissionais de saúde poderão proceder a uma análise mais exata e pormenorizada de dados de saúde complexos dos pacientes, ainda antes de adoecerem, e indicar tratamentos preventivos personalizados[65]. No contexto do envelhecimento da população europeia, a IA e a robótica podem ser ferramentas valiosas para prestar assistência aos cuidadores e apoiar os cuidados a idosos[66], bem como para monitorizar o estado de saúde dos doentes em tempo real e salvar vidas, desse modo[67].

126) A IA de confiança também pode prestar assistência numa escala mais alargada. Por exemplo, pode examinar e identificar tendências gerais no setor dos cuidados e tratamentos de saúde[68], levando a uma deteção mais precoce das doenças, a um desenvolvimento mais eficiente de medicamentos, a tratamentos mais direcionados[69] e, em última instâncias, a que mais vidas sejam salvas.

c. Educação de qualidade e transformação digital

127) As novas mutações tecnológicas, económicas e ambientais exigem que a sociedade se torne mais proativa. Os governos, os líderes da indústria, as instituições de ensino e os sindicatos têm a responsabilidade de trazer os cidadãos para a nova era digital, assegurando que possuem as competências adequadas para preencher os postos de trabalho futuros. As tecnologias de IA de confiança podem ajudar a prever com maior exatidão os postos de trabalho e as profissões que serão afetados pela tecnologia, as novas funções que serão criadas e as competências que serão necessárias. Tal poderá ajudar os governos, os sindicatos e a indústria a planear a (re)qualificação dos trabalhadores. Também poderá dar aos cidadãos que receiem o despedimento um caminho para desenvolverem novas funções.

128) Além disso, a IA pode ser uma ferramenta importante para combater as desigualdades educativas e criar programas de ensino personalizados e adaptáveis, que possam ajudar todas as pessoas a adquirirem novas

---

63 O projeto europeu UP-Drive, por exemplo, visa abordar os referidos desafios associados aos transportes fornecendo contributos que permitam a automatização gradual dos veículos e a colaboração entre si, facilitando um sistema de transporte mais inclusivo e mais acessível — https://up-drive.eu/.

64 Ver, por exemplo, o projeto REVOLVER (Repeated Cancer Evolution): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, ou o projeto Murab, que realiza biópsias mais precisas e visa diagnosticar o cancro e outras doenças mais rapidamente: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 Ver, por exemplo, o projeto Live INCITE: www.karolinska.se/en/live-incite. Este consórcio de adquirentes de cuidados de saúde desafia a indústria a desenvolver soluções inteligentes de IA e de outras TIC que permitam intervenções em termos de estilo de vida no processo perioperatório. O objetivo diz respeito a soluções de saúde em linha novas e inovadoras, que possam influenciar de forma personalizada os doentes para que tomem, antes e após a cirurgia, as medidas necessárias no seu estilo de vida para otimizar os resultados dos cuidados de saúde.

66 O projeto CARESSES, financiado pela UE, diz respeito a robôs destinados a cuidar de idosos, centrando-se na sua sensibilidade cultural: adaptam a sua forma de agir e de falar para respeitar a cultura e os hábitos do idoso a quem prestam assistência: http://caressesrobot.org/en/project/. Ver também a aplicação de IA «Alfred», um assistente virtual que ajuda os idosos a manterem-se ativos: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Além disso, o projeto EMPATTICS [EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems (capacitar os doentes para uma melhor informação e melhoria dos sistemas de comunicação)] investigará e definirá o modo como os profissionais de saúde e os doentes utilizam as TIC, incluindo os sistemas de IA, para planear as intervenções com os doentes e monitorizar a progressão do seu estado de saúde física e mental: www.empattics.eu.

67 Ver, por exemplo o MyHealth Avatar (www.myhealthavatar.eu), que oferece uma representação digital do estado de saúde dos doentes. O projeto de investigação lançou uma aplicação e uma plataforma em linha que recolhe e dá acesso a informações digitais a longo prazo sobre o estado de saúde dos doentes. Essa aplicação assume a forma de um companheiro («avatar») no domínio da saúde ao longo da vida. O MyHealthAvatar também prevê os riscos de AVC, diabetes, doenças cardiovasculares e hipertensão.

68 Ver, por exemplo, o projeto ENRICHME (www.enrichme.eu), que trata do declínio progressivo da capacidade cognitiva da população idosa. Uma plataforma integrada para a assistência à autonomia no domicílio e um robô de serviço móvel para monitorização e interação a longo prazo ajudarão os idosos a manterem-se independentes e ativos durante mais tempo.

69 Ver, por exemplo, a utilização da IA pela Sophia Genetics, que maximiza o valor dos dados genómicos e radiómicos através da inferência estatística, do reconhecimento de padrões e da aprendizagem automática: https://www.sophiagenetics.com/home.html.

qualificações, capacidades e competências consentâneas com a sua própria capacidade de aprendizagem[70]. Poderá aumentar tanto a rapidez da aprendizagem como a qualidade do ensino — desde o ensino básico até à universidade.

2. Exemplos de preocupações críticas suscitadas pela IA

129) Uma preocupação crítica no domínio da IA surge quando uma das componentes da IA de confiança é violada. Muitas das preocupações abaixo enumeradas já estarão abrangidas pelos requisitos legais existentes, que são obrigatórios, pelo que têm de ser cumpridos. Todavia, mesmo em circunstâncias em que o cumprimento dos requisitos legais foi demonstrado, estes podem não abranger todas as preocupações de ordem ética que poderão surgir. Dado que a nossa compreensão da adequação das regras e dos princípios éticos evolui invariavelmente e pode mudar ao longo do tempo, a seguinte lista não exaustiva de preocupações pode ser abreviada, alargada, editada ou atualizada no futuro.

a. Identificação e localização de pessoas com a IA

130) A IA permite a identificação cada vez mais eficiente de indivíduos por entidades públicas e privadas. São exemplos notáveis de uma tecnologia expansível de identificação baseada na IA o reconhecimento facial e outros métodos de identificação involuntários que utilizam dados biométricos (ou seja, deteção de mentiras, avaliação da personalidade através de microexpressões e reconhecimento automático de voz). A identificação de pessoas é, por vezes, um resultado desejável, consentâneo com os princípios éticos (p. ex., na deteção de fraudes, branqueamento de capitais ou financiamento do terrorismo). No entanto, a identificação automática suscita fortes preocupações de natureza ética e jurídica, uma vez que pode ter um impacto inesperado a muitos níveis psicológicos e socioculturais. É necessário fazer uma utilização proporcionada das técnicas de controlo da IA para defender a autonomia dos cidadãos europeus. Para alcançar uma IA de confiança, será essencial definir claramente se, quando e como a IA pode ser utilizada na identificação automática de pessoas e diferenciar a identificação de uma pessoa da sua localização e seguimento, bem como distinguir a vigilância direcionada da vigilância em massa. A aplicação dessas tecnologias deve estar claramente justificada pela legislação existente[71]. Caso a base jurídica de tal atividade seja o «consentimento», devem desenvolver-se meios práticos[72] que permitam dar um consentimento significativo e verificado a ser objeto de identificação automática pela IA ou por tecnologias equivalentes. O mesmo se aplica à utilização de dados pessoais «anónimos» que possam ser repersonalizados.

b. Sistemas de IA secretos

131) Os seres humanos devem saber sempre se estão a interagir diretamente com outro ser humano ou com uma máquina e é aos profissionais no domínio da IA que compete assegurar que esse conhecimento lhes é facultado de forma fiável. Por conseguinte, os profissionais no domínio da IA devem assegurar que os seres humanos são informados — ou que podem inquirir e validar o facto — de que estão a interagir com um sistema de IA (p. ex., através de declarações de exoneração de responsabilidade claras e transparentes). Note-se que existem casos-limite que complicam a questão (p. ex. a voz de um ser humano filtrada por IA). Importa recordar que a confusão entre seres humanos e máquinas pode ter múltiplas consequências, como a

[70] Ver, por exemplo, o projeto MaTHiSiS, que visa fornecer uma solução para a aprendizagem baseada nos afetos num ambiente de aprendizagem confortável, compreendendo dispositivos tecnológicos e algoritmos topo de gama: http://mathisis-project.eu/. Ver também a Watson Classroom da IBM ou a plataforma da Century Tech.

[71] Recorde-se, a este respeito, o artigo 6.º do RGPD que dispõe, nomeadamente, que o tratamento só é lícito se tiver uma base jurídica válida.

[72] Como mostram os atuais mecanismos para dar um consentimento esclarecido na Internet, os consumidores dão normalmente o seu consentimento sem uma análise significativa. Por conseguinte, dificilmente poderão ser classificados como práticos.

criação de laços, a influência ou a redução do valor da condição humana[73]. O desenvolvimento de robôs semelhantes a seres humanos[74] deverá, por isso, ser objeto de uma avaliação ética cuidadosa.

c. Classificação dos cidadãos assente na IA em violação dos direitos fundamentais

132) As sociedades devem esforçar-se por proteger a liberdade e a autonomia de todos os cidadãos. Qualquer forma de classificação dos cidadãos pode levar à perda desta autonomia e pôr em risco o princípio da não discriminação. A classificação só deve ser utilizada se existir uma justificação clara para tal e se as medidas forem proporcionadas e equitativas. A classificação normativa dos cidadãos (avaliação geral da «personalidade moral» ou da «integridade ética») em *todos* os aspetos e em larga escala por autoridades públicas ou intervenientes privados põe em risco estes valores, em especial quando não é utilizada em conformidade com os direitos fundamentais e quando é utilizada de forma desproporcionada e sem uma finalidade legítima, definida e comunicada.

133) Atualmente, a classificação dos cidadãos — em maior ou menor escala — já é muitas vezes utilizada em classificações puramente descritivas e em domínios específicos (p. ex., sistemas de ensino, aprendizagem eletrónica e cartas de condução). Mesmo nessas aplicações mais restritas, deve ser disponibilizado aos cidadãos um procedimento totalmente transparente, incluindo informações sobre o processo, a finalidade e a metodologia da classificação. Note-se que a transparência não pode evitar a não discriminação nem garantir a equidade, e que não é uma panaceia para o problema da classificação. Idealmente, deve dar-se a possibilidade de autoexclusão do mecanismo de classificação, sempre que possível sem prejuízo — caso contrário, devem ser facultados mecanismos de contestação e retificação das classificações. Este aspeto é particularmente importante nas situações em que existe uma assimetria de poder entre as partes. Essas opções de autoexclusão devem ser asseguradas na conceção da tecnologia nas situações em que tal seja necessário para garantir a conformidade com os direitos fundamentais e respeitar as regras de uma sociedade democrática.

d. Sistemas de armas letais autónomas (SALA)

134) Atualmente, um número desconhecido de países e indústrias estão a investigar e a desenvolver sistemas de armas letais autónomas, que vão desde mísseis capazes de selecionar os alvos até máquinas com aprendizagem que possuem competências cognitivas para decidir quem, quando e onde combater sem intervenção humana. Este facto suscita preocupações éticas de fundo, como o facto de poder levar a uma corrida não controlável ao armamento a um nível sem precedentes na história, e criar contextos militares em que o controlo humano é quase totalmente posto de parte e os riscos de mau funcionamento não são acautelados. O Parlamento Europeu apelou para que se adote urgentemente uma posição comum, juridicamente vinculativa, sobre as questões éticas e jurídicas relativas ao controlo humano, à supervisão, à responsabilização e à aplicação do direito internacional, humanitário e em matéria de direitos humanos, e das estratégias militares[75]. Recordando o objetivo da União Europeia de promover a paz, consagrado no artigo 3.º do Tratado da União Europeia, concordamos e apoiamos a Resolução do Parlamento Europeu de 12 de setembro de 2018 e todos os esforços conexos no domínio dos sistemas de armas letais autónomas.

e. Eventuais preocupações a longo prazo

135) O desenvolvimento de IA ainda é realizado setorialmente e exige que cientistas e engenheiros humanos com a devida formação especifiquem os seus objetivos de forma precisa. No entanto, fazendo extrapolações para um futuro mais distante, é possível formular algumas hipóteses sobre determinadas preocupações críticas a longo

[73] Madary & Metzinger (2016). «Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology.» Frontiers in Robotics and AI, 3(3).

[74] O mesmo se aplica a avatares baseados na IA.

[75] Resolução do Parlamento Europeu 2018/2752(RSP).

prazo[76]. Uma abordagem baseada nos riscos sugere que estas preocupações devem continuar a ser tidas em consideração atendendo a eventuais factos que não sabemos desconhecer e a «cisnes negros»[77]. O elevado impacto destas preocupações, conjugado com a atual incerteza quanto à evolução que poderão ter, exige que estes temas sejam regularmente avaliados.

D. CONCLUSÃO

136) O presente documento constitui as orientações éticas no domínio da IA elaboradas pelo grupo de peritos de alto nível sobre a inteligência artificial (GPAN IA).

137) Reconhecemos o impacto positivo que os sistemas de IA já têm e continuarão a ter, tanto a nível comercial como societal. No entanto, estamos igualmente interessados em garantir que os riscos e outros impactos negativos associados a estas tecnologias são geridos de forma adequada e proporcionada à luz da aplicação de IA. A IA é uma tecnologia simultaneamente transformadora e disruptiva, e a sua evolução nos últimos anos foi facilitada pela disponibilidade de enormes quantidades de dados digitais, pelos grandes avanços tecnológicos em termos de capacidade de computação e de capacidade de armazenamento, bem como pela significativa inovação científica e de engenharia nos métodos e nas ferramentas de IA. Os sistemas de IA continuarão a influenciar a sociedade e os cidadãos de formas que ainda não conseguimos imaginar.

138) Neste contexto, é importante construir sistemas de IA que sejam fiáveis, uma vez que os seres humanos só conseguirão confiar e tirar pleno proveito dos benefícios da IA quando esta tecnologia, incluindo os processos e as pessoas que lhe estão subjacentes, for digna de confiança. Por conseguinte, ao elaborar as presentes orientações, alcançar essa IA de confiança foi a nossa ambição fundamental.

139) Uma IA de confiança tem três componentes: 1) deve ser Legal, garantindo o respeito de toda a legislação e regulamentação aplicáveis; 2) deve ser Ética, garantindo a observância de princípios e valores éticos; e 3) deve ser Sólida, tanto do ponto de vista técnico como do ponto de vista social, uma vez que, mesmo com boas intenções, os sistemas de IA podem causar danos não intencionais. Cada uma destas componentes é necessária, mas não suficiente, para alcançar uma IA de confiança. Idealmente, as três componentes funcionam em harmonia, sobrepondo-se na sua ação. Quando surgem conflitos, devemos procurar harmonizá-las.

140) No capítulo I, enunciámos os direitos fundamentais e um conjunto correspondente de princípios éticos que são cruciais no contexto da IA. No capítulo II, enumerámos sete requisitos essenciais que os sistemas de IA devem satisfazer para concretizar uma IA de confiança. Propusemos métodos técnicos e não técnicos que podem contribuir para a sua aplicação. Por último, apresentámos no capítulo III uma lista de avaliação de uma IA de confiança, que pode ajudar a operacionalizar esses sete requisitos. Numa secção final, demos exemplos de oportunidades benéficas e preocupações críticas suscitadas pelos sistemas de IA, com base nos quais esperamos estimular um debate mais aprofundado.

141) A Europa tem um ponto de vista único baseado na sua determinação em colocar os cidadãos no centro dos seus esforços. Esta determinação está inscrita no próprio ADN da União Europeia através dos Tratados em que se alicerça. O presente documento faz parte de uma visão que promove uma IA de confiança, que consideramos dever ser a base sobre a qual a Europa poderá construir uma posição de liderança em matéria de sistemas de IA avançados e inovadores. Esta visão ambiciosa contribuirá para assegurar o desenvolvimento humano dos cidadãos europeus, tanto a nível individual como coletivamente. O nosso objetivo é criar uma

[76] Embora algumas pessoas considerem que a inteligência artificial geral, a consciência artificial, os agentes de moral artificial, a superinteligência ou a IA transformadora podem ser exemplos dessas preocupações a longo prazo (atualmente não existentes), muitas outras consideram que eles são irrealistas.

[77] Um «cisne negro» é um acontecimento muito raro, mas com grande impacto — tão raro que pode não ter sido observado. Por conseguinte, a probabilidade da sua ocorrência normalmente só pode ser estimada com uma elevada incerteza.

cultura de «IA de confiança para a Europa», mediante a qual os benefícios da IA possam ser usufruídos por todos de uma forma que garanta o respeito dos nossos valores fundamentais: os direitos fundamentais, a democracia e o Estado de direito.

GLOSSÁRIO

142) O presente glossário diz respeito às orientações e destina-se a ajudar a compreender os termos utilizados neste documento.

Inteligência artificial ou sistemas de IA

143) Os sistemas de inteligência artificial (IA) são sistemas de software (e eventualmente também de hardware) concebidos por seres humanos[78], que, tendo recebido um objetivo complexo, atuam na dimensão física ou digital percecionando o seu ambiente mediante a aquisição de dados, interpretando os dados estruturados ou não estruturados recolhidos, raciocinando sobre o conhecimento ou processando as informações resultantes desses dados e decidindo as melhores ações a adotar para atingir o objetivo estabelecido. Os sistemas de IA podem utilizar regras simbólicas ou aprender um modelo numérico, bem como adaptar o seu comportamento mediante uma análise do modo como o ambiente foi afetado pelas suas ações anteriores.

144) Enquanto disciplina científica, a IA inclui diversas abordagens e técnicas, tais como a aprendizagem automática (de que a aprendizagem profunda e a aprendizagem por reforço são exemplos específicos), o raciocínio automático (que inclui o planeamento, a programação, a representação do conhecimento e o raciocínio, a pesquisa e a otimização) e a robótica (que inclui o controlo, a perceção, os sensores e atuadores, bem como a integração de todas as outras técnicas em sistemas ciberfísicos).

145) Um documento separado elaborado pelo GPAN IA e que explicita a definição de *sistemas de IA* utilizada no presente documento é publicado em paralelo, com o título «Uma definição de IA: Principais capacidades e disciplinas científicas».

Profissionais no domínio da IA

146) Por profissionais no domínio da IA entende-se todos os indivíduos ou organizações que desenvolvem (incluindo investigação, conceção ou fornecimento de dados para o desenvolvimento), implantam (incluindo a aplicação) ou utilizam sistemas de IA, excluindo aqueles que os utilizam na qualidade de utilizador final ou consumidor.

Ciclo de vida do sistema de IA

147) O ciclo de vida do sistema de IA engloba as suas fases de desenvolvimento (incluindo investigação, conceção, fornecimento de dados e ensaios limitados), implantação (incluindo aplicação) e utilização.

Auditabilidade

148) A auditabilidade refere-se à capacidade de um sistema de IA se sujeitar à avaliação dos seus algoritmos, dados e processos de conceção. Constitui um dos sete requisitos que um sistema de IA de confiança deve cumprir. Tal não implica necessariamente que as informações sobre os modelos de negócios e a propriedade intelectual relacionadas com o sistema de IA tenham de estar sempre publicamente disponíveis. Assegurar mecanismos de rastreabilidade e de registo cronológico desde o início da fase de conceção do sistema de IA pode contribuir para possibilitar a auditabilidade do sistema.

Enviesamento

149) Entende-se por enviesamento uma tendência parcial a favor ou contra uma pessoa, um objeto ou uma posição. Os enviesamentos podem surgir de muitas formas nos sistemas de IA. Por exemplo, nos sistemas de IA baseados em dados, como os produzidos por via da aprendizagem automática, o enviesamento na recolha de dados e na fase de treino pode levar um sistema de IA que apresenta enviesamentos. Na IA baseada na lógica, como os sistemas baseados em regras, podem surgir enviesamentos devido à forma como um engenheiro do conhecimento entenda as regras aplicáveis num determinado contexto. Também podem surgir

[78] Os seres humanos concebem os sistemas de IA diretamente, mas também podem utilizar técnicas de IA para otimizar a sua conceção.

enviesamentos devido à aprendizagem em linha e à adaptação através da interação. Podem ainda surgir através da personalização, que visa apresentar aos utilizadores recomendações ou fluxos de informações adaptadas aos seus gostos. Não estão necessariamente relacionados com preconceitos humanos ou uma recolha de dados baseada no ser humano. Podem ser suscitados, por exemplo, pelos contextos limitados em que um sistema é utilizado, não havendo nesse caso oportunidades de generalização para outros contextos. O enviesamento pode ser bom ou mau, intencional ou não intencional. Em alguns casos, o enviesamento pode causar resultados discriminatórios e/ou injustos, designados no presente documento por enviesamentos injustos.

Ética

150) A ética é uma disciplina académica que constitui um subdomínio da filosofia. Em termos gerais, trata de questões como «O que é uma boa ação?», «Qual é o valor de uma vida humana?», «O que é a justiça?» ou «O que é uma boa vida?». Na ética académica, há quatro grandes domínios de investigação: i) metaética, que diz sobretudo respeito ao significado e à referência das frases normativas, e à questão de saber como os seus valores de verdade podem ser determinados (caso existam), ii) ética normativa, um meio prático para determinar a orientação moral a seguir, mediante um exame das normas de boa e má ação e da atribuição de um valor a ações específicas, iii) ética descritiva, que visa fazer uma investigação empírica do comportamento e das convicções morais das pessoas, iv) ética aplicada, respeitante ao que somos obrigados (ou autorizados) a fazer numa situação específica (muitas vezes historicamente nova) ou num determinado domínio (muitas vezes sem precedentes históricos) de possibilidades de ação. A ética aplicada trata de situações da vida real, em que as decisões têm de ser tomadas sob pressão do tempo e muitas vezes com uma racionalidade limitada. A ética da IA é geralmente encarada como um exemplo de ética aplicada e centra-se nas questões normativas suscitadas pela conceção, pelo desenvolvimento, pela implantação e pela utilização da inteligência artificial.

151) No âmbito dos debates éticos, os termos «moral» e «ético» são frequentemente utilizados. O termo «moral» refere-se aos padrões de comportamento concretos e factuais, aos costumes e convenções que podem ser encontrados em culturas, grupos ou indivíduos específicos num determinado momento. O termo «ético» refere-se a uma avaliação valorativa das ações concretas e dos comportamentos de uma perspetiva sistemática e académica.

IA ética

152) No presente documento, o termo «IA ética» é utilizado para indicar o desenvolvimento, a implantação e a utilização de IA que assegure o cumprimento das normas éticas, incluindo direitos fundamentais como direitos morais especiais, princípios éticos e valores fundamentais com eles relacionados. É o segundo dos três elementos essenciais necessários para alcançar uma IA de confiança.

IA centrada no ser humano

153) A abordagem à IA centrada no ser humano procura garantir que os valores humanos estejam no centro do desenvolvimento, da implantação, da utilização e do controlo dos sistemas de IA, assegurando o respeito dos direitos fundamentais, nomeadamente os direitos consagrados nos Tratados da União Europeia e na Carta dos Direitos Fundamentais da União Europeia, os quais têm como referência uma base comum alicerçada no respeito da dignidade humana, na qual o ser humano goza de um estatuto moral único e inalienável. Isto implica também que o ambiente natural e os outros seres vivos que fazem parte do ecossistema humano sejam tomados em consideração, bem como a adoção de uma abordagem sustentável que permita que as gerações futuras realizem o seu potencial.

Red Teaming

154) Entende-se por *red teaming* a prática em que uma «equipa vermelha» ou grupo independente desafia uma organização a melhorar a sua eficácia assumindo um papel ou ponto de vista antagónico. É sobretudo utilizada para ajudar a identificar e solucionar eventuais vulnerabilidades de segurança.

Reprodutibilidade

155) A reprodutibilidade descreve se uma experiência de IA apresenta o mesmo comportamento quando repetida nas mesmas condições.

IA sólida

156) A solidez de um sistema de IA abrange tanto a sua solidez técnica (adequada num determinado contexto, como o domínio de aplicação ou a fase do ciclo de vida) como a sua solidez do ponto de vista social (assegurando que o sistema de IA tem devidamente em conta o contexto e o ambiente em que o sistema opera). Este aspeto é crucial para garantir que, mesmo com boas intenções, não se podem produzir danos não intencionais. A solidez é a terceira das três componentes necessárias para alcançar uma IA de confiança.

Partes interessadas

157) Entendemos por partes interessadas todos os que investigam, desenvolvem, concebem, implantam ou utilizam a IA, bem como aqueles que são (direta ou indiretamente) afetados pela IA — incluindo, entre outros, empresas, organizações, investigadores, serviços públicos, instituições, organizações da sociedade civil, governos, reguladores, parceiros sociais, indivíduos, cidadãos, trabalhadores e consumidores.

Rastreabilidade

158) A rastreabilidade de um sistema de IA refere-se à capacidade de acompanhar os dados do sistema e os processos de desenvolvimento e implantação do mesmo, normalmente por meio de uma identificação registada documentada.

Confiança

159) Extraímos a seguinte definição da literatura: «Entende-se por confiança: 1) um conjunto de convicções específicas relacionadas com a benevolência, a competência, a integridade e a previsibilidade (convicções de confiança); 2) a disponibilidade de uma parte para depender de outros numa situação de risco (intenção de confiança); ou 3) a combinação destes elementos»[79]. Embora a «confiança» não seja uma propriedade atribuída a máquinas, o presente documento pretende salientar a importância de se poder confiar não só no facto de os sistemas de IA cumprirem a lei, respeitarem os princípios éticos e serem sólidos, mas também de se poder depositar tal confiança em todas as pessoas e todos os processos envolvidos no ciclo de vida do sistema de IA.

IA de confiança

160) Uma IA de confiança tem três componentes: 1) deve ser Legal, garantindo o respeito de toda a legislação e regulamentação aplicáveis; 2) deve ser Ética, demonstrando respeito e garantindo a observância de princípios e valores éticos; e 3) deve ser Sólida, tanto do ponto de vista técnico como do ponto de vista social, uma vez que, mesmo com boas intenções, os sistemas de IA podem causar danos não intencionais. Uma IA de confiança diz respeito não só à fiabilidade do próprio sistema de IA, mas também à fiabilidade de todos os processos e intervenientes que fazem parte do ciclo de vida do sistema.

[79] Siau, K., Wang, W. (2018), «Building Trust in Artificial Intelligence, Machine Learning, and Robotics», *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. 47–53.

Pessoas e grupos vulneráveis

161) Não existe qualquer definição jurídica comummente aceite ou consensual de pessoas vulneráveis, devido à sua heterogeneidade. O que constitui uma pessoa ou um grupo vulnerável depende frequentemente do contexto específico. Este pode ser influenciado por acontecimentos temporários da vida (p. ex., infância ou doença), fatores de mercado (p. ex., assimetrias de informação ou de poder de mercado), fatores económicos (p. ex., a pobreza), fatores ligados à identidade individual (p. ex., o género, a religião ou a cultura) ou por outros fatores. A Carta dos Direitos Fundamentais da União Europeia inclui, no artigo 21.º, relativo à não discriminação, os seguintes motivos de discriminação, que podem constituir um ponto de referência, nomeadamente: o sexo, a raça, a cor, a origem étnica ou social, as características genéticas, a língua, a religião ou as convicções, as opiniões políticas ou outras, a pertença a uma minoria nacional, a riqueza, o nascimento, a deficiência, a idade ou a orientação sexual. Outras disposições legais abordam os direitos de grupos específicos, para lá dos acima referidos. Qualquer lista deste tipo não é exaustiva e pode ser alterada ao longo do tempo. Um grupo vulnerável é um grupo de pessoas que partilham uma ou mais características de vulnerabilidade.

O presente documento foi elaborado pelos membros do grupo de peritos de alto nível sobre a inteligência artificial

a seguir referidos por ordem alfabética do sobrenome

Pekka Ala-Pietilä, presidente do GPAN IA
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann — Correlator
Zalando
Mária Bieliková
Universidade de Tecnologia Eslovaca, Bratislava
Cecilia Bonefeld-Dahl — Correlatora
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
Iniciativa para a Ética dos Sistemas Inteligentes/Autónomos da IEEE e Universidade Sorbonne
Mark Coeckelbergh
Universidade de Viena
Virginia Dignum — Correlatora
Universidade de Umeå
Luciano Floridi
Universidade de Oxford
Jean-Francois Gagné — Correlator
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agência dos Direitos Fundamentais da União Europeia
Sami Haddadin
Instituto de Robótica e Inteligência Automática da Universidade Técnica de Munique
Gry Hasselbalch
The thinkdotank DataEthics e Universidade de Copenhaga
Fredrik Heintz
Universidade de Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Universidade de Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wien
Robert Kroplewski
Advogado e consultor do Governo polaco
Elisabeth Ling
Pierre Lucas
Orgalim — Europe's technology industries
Ieva Martinkenaite
Telenor
Thomas Metzinger — Correlator
JGU Mainz & Associação das Universidades Europeias
Catelijne Muller
ALLAI Netherlands & CESE
Markus Noga
SAP
Barry O'Sullivan, vice-presidente do GPAN IA
University College Cork
Ursula Pachl
BEUC
Nicolas Petit — Correlator
Universidade de Liège
Christoph Peylo
Bosch
Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
Colégio da Europa e Centro de Estudos de Política Europeia
Francesca Rossi
IBM
Cristina San José
Federação Bancária Europeia
George Sharkov
Digital SME Alliance
Philipp Slusallek
Centro Alemão de Investigação da Inteligência Artificial (DFKI)
Françoise Soulié Fogelman
Consultora de IA
Saskia Steinacker — Correlatora
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe — Correlatora
TU Delft
Thiébaut Weber
CES
Cecile Wendling
AXA
Karen Yeung — Correlatora
Universidade de Birmingham

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe e Karen Yeung foram relatores do presente documento.

Pekka Ala-Pietilä é presidente do GPAN IA. Barry O'Sullivan é vice-presidente e responsável pela coordenação do segundo documento a elaborar pelo GPAN IA. Nozha Boujemaa, vice-presidente até 1 de fevereiro de 2019, coordenou a elaboração deste primeiro documento e contribuiu igualmente para o seu conteúdo.

Nathalie Smuha prestou apoio editorial.