Comissão Europeia

GRUPO DE PERITOS DE ALTO NÍVEL SOBRE A INTELIGÊNCIA ARTIFICIAL

PROJETO
DIRETRIZES DEONTOLÓGICAS PARA UMA IA DIGNA DE CONFIANÇA

RESUMO

Documento de trabalho para consulta das partes interessadas

Bruxelas, 18 de dezembro de 2018

PROJETO DE ORIENTAÇÕES ÉTICAS

PARA UMA IA DIGNA DE CONFIANÇA

Grupo de peritos de alto nível sobre a inteligência artificial
Projeto de orientações éticas para uma IA digna de confiança

Comissão Europeia
Direção-Geral da Comunicação

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

Comissão Europeia
B-1049 Bruxelas

Documento divulgado ao público em 18 de dezembro de 2018, em inglês.

Este documento de trabalho foi elaborado pelo GPAN IA, sem prejuízo da posição individual dos seus membros relativamente a pontos específicos e sem prejuízo da versão final do documento. Este documento será ainda desenvolvido, devendo ser apresentada uma versão final em março de 2019, na sequência da consulta das partes interessadas através da Aliança Europeia da Inteligência Artificial.

 Este é um projeto do primeiro resultado do GPAN IA. Uma versão final será apresentada à Comissão em março de 2019. Uma versão final do segundo resultado — as recomendações sobre a política em matéria de IA e de investimento — será apresentada em meados de 2019.

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).

RESUMO

O presente documento de trabalho constitui um projeto das orientações éticas no domínio da IA elaboradas pelo Grupo de peritos de alto nível sobre a inteligência artificial da Comissão Europeia (GPAN IA), cuja versão final está prevista para março de 2019.

A inteligência artificial (IA) é uma das forças mais transformadoras do nosso tempo e vai provocar alterações no tecido social. Constitui uma excelente oportunidade para aumentar a prosperidade e o crescimento, que a Europa deve procurar alcançar. Na última década, foram realizados grandes progressos devido à disponibilidade de enormes quantidades de dados digitais, a potentes arquiteturas de computação e a avanços em técnicas de IA, como a aprendizagem automática. Importantes desenvolvimentos potenciados pela IA em veículos autónomos, nos cuidados de saúde, em robôs domésticos/de serviço, na educação ou na cibersegurança estão a melhorar diariamente a qualidade das nossas vidas. Além disso, a IA é fundamental para fazer face a muitos dos grandes desafios que o mundo enfrenta, em áreas como a saúde e o bem-estar, as alterações climáticas, os sistemas jurídicos e democráticos fiáveis e outros que são expressos nos Objetivos de Desenvolvimento Sustentável das Nações Unidas.

Embora tenha a capacidade de gerar enormes benefícios para os indivíduos e para a sociedade, a IA também dá origem a certos riscos que devem ser devidamente geridos. Dado que, de um modo geral, os benefícios da IA são superiores aos seus riscos, temos de assegurar que seguimos o caminho que nos permita **maximizar os benefícios da IA, minimizando ao mesmo tempo os seus riscos**. Para garantir que nos mantemos no bom caminho, **é necessária uma abordagem centrada no ser humano,** obrigando-nos a ter em mente que o desenvolvimento e a utilização da IA não devem ser encarados como um meio em si mesmo, mas como o objetivo de aumentar o bem-estar humano. Uma **IA de confiança será o nosso princípio orientador**, uma vez que os seres humanos só poderão usufruir plenamente e de forma confiante dos benefícios da IA se puderem confiar na tecnologia.

Uma IA de confiança tem **duas componentes:** 1) deve respeitar os direitos fundamentais, a regulamentação aplicável e os princípios e valores fundamentais, garantindo um «**objetivo ético**» e 2) deve ser **tecnicamente sólida** e fiável, uma vez que, mesmo com boas intenções, a falta de domínio tecnológico pode causar danos não intencionais.

Por conseguinte, as presentes orientações estabelecem um **quadro para uma IA de confiança:**

- O **capítulo I** trata de **garantir o objetivo ético da IA**, estabelecendo os direitos, princípios e valores fundamentais que esta deve respeitar.
- A partir desses princípios, o **capítulo II** fornece **orientações sobre a realização** de uma IA de confiança, abordando simultaneamente o objetivo ético e a solidez técnica. Para tal, é necessário enumerar os requisitos de uma IA de confiança e oferecer uma panorâmica dos métodos técnicos e não técnicos que podem ser utilizados para a sua aplicação.
- O **capítulo III operacionaliza** subsequentemente os requisitos, fornecendo uma lista de avaliação concreta, mas não exaustiva, de uma IA de confiança. Esta lista é, em seguida, adaptada a casos específicos de utilização.

Em contraste com outros documentos relacionados com a IA ética, as orientações não pretendem fornecer mais uma lista de valores e princípios fundamentais para a IA, mas sim fornecer orientações sobre a execução concreta e respetiva operacionalização em sistemas de IA. Essas orientações são fornecidas em três níveis de abstração, do nível mais elevado de abstração no capítulo I (direitos fundamentais, princípios e valores), até ao nível mais concreto no capítulo III (lista de avaliação).

As orientações destinam-se a todas as **partes interessadas relevantes que desenvolvem, aplicam ou utilizam a IA**, englobando empresas, organizações, investigadores, serviços públicos, instituições, pessoas singulares ou outras entidades. Na versão final das presentes orientações, será apresentado um

mecanismo que permita às partes interessadas subscrever voluntariamente as referidas orientações.

Mais importante ainda, as presentes orientações não se destinam a substituir qualquer forma de decisão política ou regulamentação (tema a ser tratado no segundo resultado do GPAN IA: recomendações sobre a política em matéria de IA e de investimento, previstas para maio de 2019), nem se destinam a dissuadir a sua introdução. Além disso, as orientações devem ser vistas como um documento dinâmico que necessita de ser regularmente atualizado, a fim de garantir que continua a ser relevante à medida que a tecnologia e o nosso conhecimento da mesma evoluem. Este documento deve, por conseguinte, constituir um ponto de partida para o debate sobre o tema **«Uma inteligência artificial de confiança na Europa»**.

Embora a Europa apenas possa transmitir a sua abordagem ética à IA se for competitiva a nível mundial, é **essencial uma abordagem ética da IA para permitir uma competitividade responsável**, uma vez que gerará confiança por parte dos utilizadores e facilitará uma maior aceitação da IA. As presentes orientações não se destinam a asfixiar a inovação no domínio da IA na Europa, mas sim a utilizar a ética como fonte de inspiração para desenvolver uma marca única de IA, que visa proteger e beneficiar tanto os indivíduos como o bem comum. Tal permitirá à Europa posicionar-se como líder no domínio da IA de topo, segura e ética. Só garantindo a fiabilidade da IA é que os cidadãos europeus poderão colher plenamente os seus benefícios.

Por último, as presentes orientações visam igualmente **promover**, para além da Europa, **a reflexão e o debate** sobre um quadro ético para a IA a **nível mundial**.

ORIENTAÇÕES DE EXECUÇÃO

Cada capítulo das orientações proporciona indicações sobre como alcançar uma IA de confiança, dirigidas a todas as partes interessadas que desenvolvem, aplicam ou utilizam a IA, e que são resumidas a seguir:

Capítulo I: Orientações fundamentais para garantir o objetivo ético:

- Assegurar que a IA é **centrada no ser humano**: A IA deve ser desenvolvida, aplicada e utilizada com um **objetivo ético**, baseada em, e refletindo os direitos fundamentais, os valores societais e os princípios éticos da *beneficência* (fazer o bem), *não-maleficiência* (não fazer o mal), *autonomia do ser humano, justiça* e *explicabilidade.* Este aspeto é crucial para o trabalho no sentido de uma **IA de confiança**.
- Basear-se em direitos fundamentais, em princípios éticos e em valores para avaliar prospetivamente os efeitos possíveis da IA nos seres humanos e no bem comum. Prestar **especial atenção** a situações que envolvam **grupos mais vulneráveis**, como crianças, pessoas com deficiência ou minorias, ou a situações com **assimetrias de poder ou de informação**, tais como entre empregadores e empregados ou entre empresas e consumidores.
- Reconhecer e estar ciente do facto de que, ao mesmo tempo que traz benefícios substanciais aos indivíduos e à sociedade, a IA também pode ter um impacto negativo. Permanecer vigilante em relação a áreas que suscitam grande preocupação.

Capítulo II: Orientações fundamentais para a realização de uma IA de confiança:

- Incorporar os **requisitos para uma IA de confiança desde a primeira fase de conceção**: Responsabilização, governação dos dados, conceção para todos, governação da autonomia da IA (supervisão humana), não-discriminação, respeito pela autonomia humana, respeito pela privacidade, solidez, segurança e transparência.
- Considerar métodos técnicos e não técnicos para assegurar a aplicação desses requisitos no sistema de IA. Além disso, ter em consideração estes requisitos ao constituir a equipa que trabalha no

sistema, ao construir o sistema propriamente dito, o ambiente de testes e as potenciais aplicações do sistema.

- Fornecer, de forma clara e proativa, **informações às partes interessadas** (clientes, trabalhadores, etc.) sobre as capacidades e as limitações do sistema de IA, permitindo-lhes criar expectativas realistas. Neste contexto, é fundamental garantir a **rastreabilidade** do sistema de IA.
- Tornar a IA de confiança **parte da cultura da organização** e fornecer informações às partes interessadas sobre a forma como a IA de confiança deve ser transposta para a conceção e utilização dos sistemas de IA. Uma IA de confiança pode também ser incluída nas cartas de deontologia ou nos códigos de conduta das organizações.
- Assegurar a participação e a **inclusão das partes interessadas** na conceção e no desenvolvimento do sistema de IA. Além disso, assegurar a **diversidade** das equipas que desenvolvem, implementam e testam o produto.
- Procurar **facilitar a auditabilidade** dos sistemas de IA, em particular em contextos ou situações críticos. Na medida do possível, conceber o sistema de modo a permitir o rastreio das decisões individuais para os vários contributos; dados, modelos previamente formados, etc. Além disso, definir os **métodos de explicação** do sistema de IA.
- Assegurar um processo específico de **responsabilização da governação**.
- Prever a **formação e a educação** e garantir que os gestores, os criadores, os utilizadores e os empregadores têm conhecimento e recebem formação em matéria de inteligência artificial de confiança.
- Estar ciente de que podem existir tensões fundamentais entre diferentes objetivos (a transparência pode conduzir a uma má utilização; a identificação e correção dos enviesamentos podem pôr em risco a proteção da privacidade). Comunicar e documentar essas soluções de compromisso.
- Promover a investigação e a inovação para melhorar o cumprimento dos requisitos da IA de confiança.

Capítulo III: Orientações fundamentais para a avaliação de uma IA de confiança:

- 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.
- Não esquecer que uma lista de avaliação **nunca será exaustiva** e que garantir uma IA de confiança não se resume a um exercício de preenchimento de quadrículas, mas se trata de um processo contínuo de identificação dos requisitos, de avaliação de soluções e de garantia de melhores resultados ao longo de todo o ciclo de vida do sistema de IA.

Estas orientações fazem parte de uma visão que engloba uma abordagem da IA centrada no ser humano, que permitirá à Europa tornar-se um líder mundial em inovação em matéria de IA ética, segura e de vanguarda. O seu objetivo é facilitar uma **«IA de confiança criada na Europa»**, que reforçará o bem-estar dos cidadãos europeus.