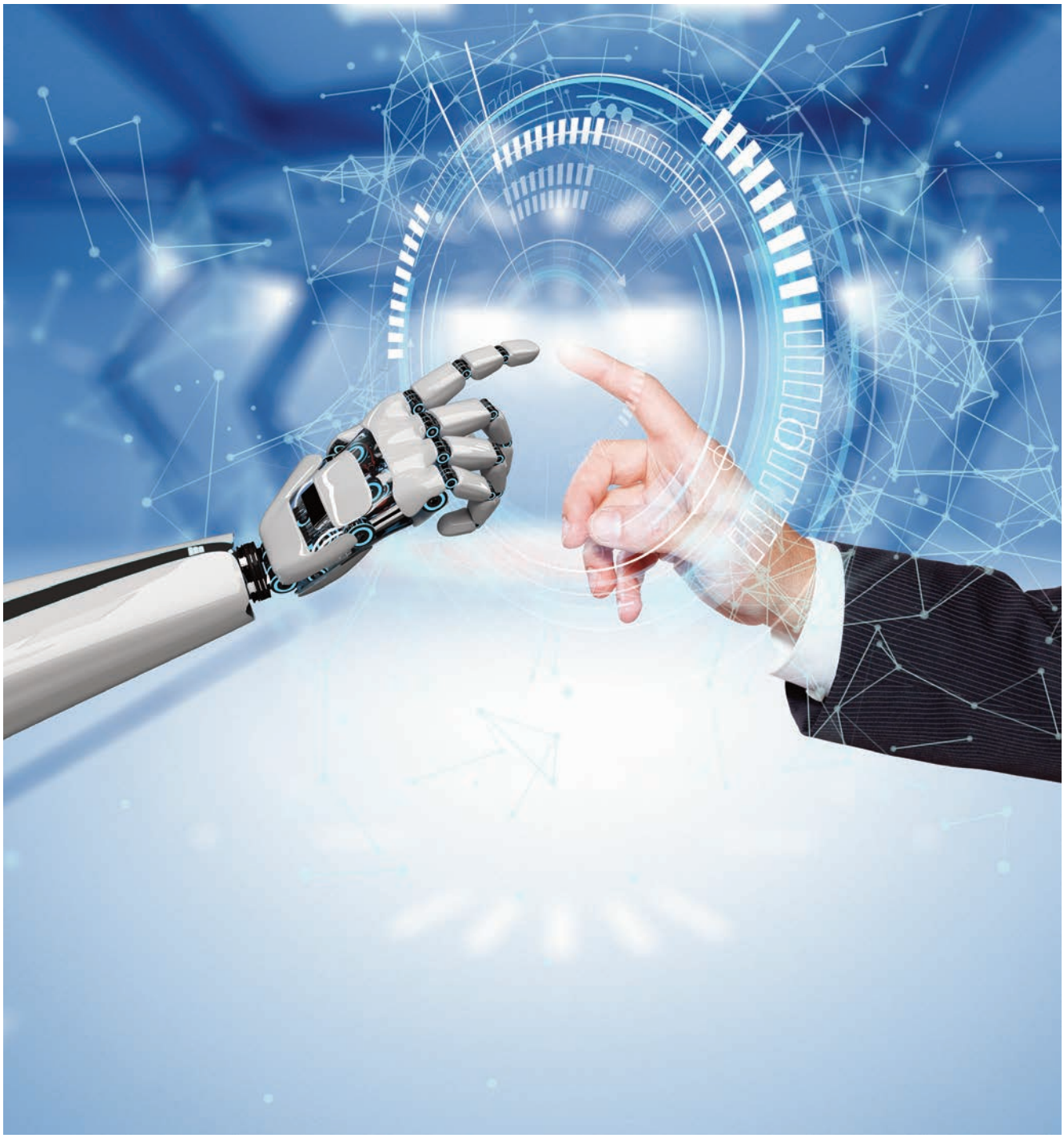
Contexto e fundamentos da XAI

"Compreender a inteligência artificial é um desafio que exige uma enorme capacidade intelectual; felizmente, temos a inteligência artificial para lidar com isso."
GPT-4[16]



Contexto

Um dos recursos mais notáveis da transformação digital é que ela está disponibilizando para todos os setores uma quantidade enorme de dados estruturados e não estruturados provenientes de vários aplicativos; por exemplo:

- Dados de varejo procedentes de ações de compra, transações e feedbacks dos clientes.
- Dados financeiros de fontes bancárias, de investimento e comerciais.
- Dados de redes sociais, incluindo análise de opiniões e análise preditiva.
- Sensores digitais de IoT (Internet das Coisas) que medem temperatura, pressão e outros dados do entorno.
- Dados de saúde, como registros médicos, diagnósticos, imagens e informações genômicas.
- Wearables, como rastreadores de atividade, sensores de saúde e smartwatches.
- Sistemas de reconhecimento de fala que permitem que as máquinas entendam e respondam à linguagem natural.
- Satélites e outros sensores espaciais que fornecem informações sobre o tempo e o clima.
- Sistemas de vigilância inteligente usando reconhecimento facial e detecção de objetos.
- Sensores de veículos autônomos, como câmeras, lidar, radar e sensores ultrassônicos.

A disponibilidade desses dados, aliada à presença de enormes recursos de armazenamento e de processamento computacional a um custo reduzido, impulsionou um maior apetite por modelagem avançada, manifestado no uso de uma ampla variedade de técnicas de machine learning e no desenvolvimento de inteligência artificial (AI) em praticamente todos os setores e âmbitos[17].

Embora haja consenso de que os modelos de AI geralmente oferecem maior poder de previsão do que os modelos tradicionais[18], teles também introduzem maior complexidade e podem ser difíceis de interpretar e explicar seus resultados.

Isso cria riscos associados ao uso desses modelos, como a falta de compreensão do modelo, a presença de vieses inadvertidos ou a dificuldade de determinar se o modelo está treinado em excesso (global ou localmente), o que pode levar a uma escassa capacidade de generalização e a possíveis erros nas decisões baseadas no modelo e, consequentemente, a uma falta de confiança no modelo.

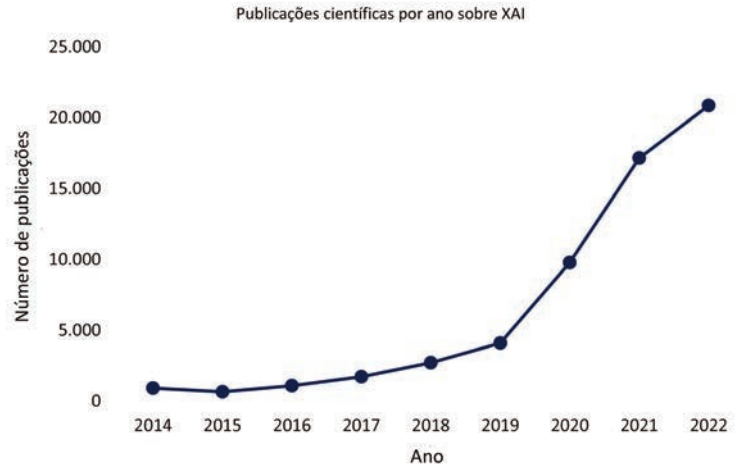
Isso levanta a questão de saber se é possível entender suficientemente bem os resultados dos algoritmos de AI, especialmente quando eles afetam decisões críticas, como diagnóstico médico, direção autônoma, detecção de fraudes e muitas outras.

[16] GPT-4, Generative Pre-Trained Transformer, uma rede neural profunda projetada pela OpenAI Foundation para executar tarefas de processamento de linguagem natural (NLP). Nesse caso, foi solicitado a ele que "apresentasse 10 citações inteligentes sobre inteligência artificial e quão difícil e necessário é ser capaz de interpretar e explicar os modelos de AI". A citação enviada foi a terceira.

[17] Embora existam diferenças, dada a falta de consenso sobre sua definição, os termos "aprendizado de máquina", "machine learning (ML)", "inteligência artificial (AI)" e "modelagem avançada" serão usados de forma intercambiável neste documento. Além disso, a abreviação "AI" será usada para "inteligência artificial", para fins de consistência com o acrônimo "XAI" (que geralmente não é traduzido), mesmo em citações de publicações em português.

[18] LeCun, Y. et al (2015). Pesquisador do Facebook AI Research e da Universidade de Nova York.

Figura 2. Número de publicações científicas por ano sobre Inteligência Artificial Explicável (XAI).



Definição

A disciplina de XAI é relativamente nova e, portanto, ainda não há uma doutrina estabelecida que padronize sua terminologia. Apesar de alguns esforços notáveis para definir os termos[19], a abordagem da XAI é heterogênea (dependendo da fonte acadêmica consultada) ou intuitiva (mais frequente na prática industrial).

De qualquer forma, para a maioria dos usos na prática, pode ser suficiente definir XAI da seguinte forma[20]:

> A inteligência artificial explicável (XAI) é o conjunto de processos e métodos que permite que os usuários humanos entendam e confiem nos resultados e produtos criados pelos algoritmos de machine learning. A XAI é usada para descrever um modelo de AI, seu impacto esperado e possíveis vieses. Ela ajuda a caracterizar a precisão, a imparcialidade, a transparência e os resultados do modelo na tomada de decisões baseada em AI. A XAI é fundamental para que uma organização crie confiança ao colocar modelos de AI em produção. A explicabilidade da AI também ajuda a organização a adotar uma abordagem responsável para o desenvolvimento da AI.

Relevância da XAI

Um aspecto sobre o qual há consenso entre acadêmicos e profissionais do setor é a crescente relevância da XAI como uma disciplina complementar à AI.

As ferramentas de análise de publicações científicas identificam mais de 77.000 artigos sobre XAI entre 2014 e 2022, e em uma tendência de aumento exponencial, com mais de 20.000 artigos somente em 2022 (Fig. 2)[21].

Além do interesse acadêmico, a atenção dada à XAI é explicada por sua capacidade de abordar uma série de preocupações do setor no uso da AI (Fig. 3), entre elas:

- **Requisitos regulatórios:** a obrigação de cumprir a regulamentação emergente sobre o uso de AI.
- **Falta de confiança:** a necessidade de criar confiança no modelo de AI e nos resultados que ele fornece entre os usuários, validadores e auditores e, em última análise, o público em geral.
- **Potencial uso indevido:** a conveniência de evitar o uso indevido dos modelos devido à falta de compreensão de como eles funcionam, o que pode resultar em custos e até mesmo em penalidades.
- **Impacto reputacional:** a prevenção de impactos reputacionais na empresa devido a preconceitos, decisões discriminatórias, uso inadequado ou simplesmente previsões errôneas do modelo.
- **Impactos sociais ou humanos:** a prevenção de impactos sociais ou humanos em usos críticos, como AI para diagnóstico de doenças médicas, decisões judiciais, identificação biométrica, polígrafos, etc.
- **Outros:** mitigação de outros riscos decorrentes da falta de compreensão do modelo, como segurança cibernética, proteção de dados, fraude, risco de modelo, etc.

Apesar de tudo isso, há casos em que os modelos de AI não precisam ser particularmente interpretáveis, porque os usos não estão regulamentados, porque não têm impactos potenciais relevantes ou simplesmente porque não precisam ser interpretados, como sistemas de recomendação automática de filmes e músicas ou algoritmos que jogam xadrez, por exemplo.

[19] Marcinkevics et al. (2020). Departamento de Ciência da Computação, ETH Zurich.
[20] IBM (2022).
[21] Dimensions (2022).

Figura 3. Preocupações do setor que a XAI está ajudando a resolver.

Regulamentação

Portanto, a XAI está se posicionando como uma disciplina de relevância crescente, o que está levando os órgãos reguladores e supervisores de diferentes jurisdições a estabelecer regulamentos e diretrizes para o uso adequado da AI, incluindo aspectos de interpretabilidade do modelo.

Nesse contexto, possivelmente as referências regulatórias mais relevantes no momento em que este artigo foi escrito são as seguintes:

1. GDPR (Parlamento Europeu)

Na Europa, o Regulamento Geral de Proteção de Dados, que entrou em vigor em 2018, estabelece o "direito a uma explicação" para os cidadãos, segundo o qual[22]:

> O titular dos dados deve ter o **direito de não estar sujeito a uma decisão,** que pode incluir uma medida, que avalie aspectos pessoais relacionados a ele, que se baseie exclusivamente no processamento automatizado e que produza efeitos legais sobre ele ou que o afete significativamente de forma semelhante, como a recusa automática de uma solicitação de crédito on-line ou de serviços de compras on-line em que não haja intervenção humana. [...]
>
> De qualquer forma, esse processamento deve estar sujeito a salvaguardas adequadas, que devem incluir informações específicas ao titular dos dados e o direito de obter intervenção humana, de expressar seu ponto de vista, de **receber uma explicação sobre a decisão tomada após essa avaliação** e de contestar a decisão.

Isso tem implicações críticas para o uso da AI e pode levar a questionamentos sobre sua viabilidade. Entretanto, nas palavras do Parlamento Europeu[23]:

> Certamente há uma tensão entre os princípios tradicionais de proteção de dados - limitação da finalidade, minimização de dados, tratamento especial de "dados sensíveis", limitação de decisões automatizadas - e a implantação total do poder da AI e do big data. Esses últimos envolvem a coleta de grandes quantidades de dados relacionados a indivíduos e suas relações sociais e seu processamento para fins que não foram totalmente determinados no momento da coleta. Entretanto, **há maneiras de interpretar, aplicar e desenvolver princípios de proteção de dados que sejam consistentes com os usos benéficos da AI e do big data.**

E isso está de acordo com o quarto princípio para o uso ético da AI estabelecido pelo Grupo de Alto Nível da Comissão Europeia sobre Inteligência Artificial[24]:

> **Explicabilidade:** os processos algorítmicos devem ser transparentes, os recursos e os objetivos dos sistemas de AI devem ser comunicados abertamente e as decisões devem ser explicadas às pessoas direta e indiretamente afetadas.

De qualquer forma, o GDPR tem impactos relevantes sobre o uso da AI, no sentido de que as empresas são legalmente obrigadas a explicar por que um modelo de AI produziu um determinado resultado, e isso tem implicações críticas para o desenho e a análise de interpretabilidade dos modelos de AI[25].

2. Artificial intelligence act (Parlamento Europeu)

O projeto de Regulamento de Inteligência Artificial ou Artificial Intelligence Act (AI Act), publicado em 2021, é uma proposta para o uso de inteligência artificial na União Europeia que visa garantir um alto nível de confiança na AI e em suas aplicações, ao mesmo tempo em que estabelece as bases para a inovação.

[22] GDPR (2018), Cons. 71.
[23] European Parliamentary Research Service (2020).
[24] Ibíd.
[25] Em alguns países europeus, o nível de conformidade desse tipo de AI (em especial os chamados Large Language Models) com a regulamentação de proteção de dados está sendo analisado e, em alguns casos, o uso de alguns desses modelos foi temporariamente proibido.

O regulamento estabelece um framework regulatório para sistemas de AI na UE e inclui requisitos de desenvolvimento ético, transparência, segurança e precisão. Ele também estabelece um sistema de governança e supervisão para sistemas de AI, bem como regras de proteção e governança de dados.

Sendo um regulamento, quando for adotado, será de aplicação direta nos 27 países da União[26], sem a necessidade de ser transposto para a legislação de cada país.

Uma de suas principais características é que ele classifica os aplicativos de AI em níveis de risco[27]:

- **Práticas proibidas,** que denotam a categoria de maior risco; esses sistemas são totalmente proibidos. Eles incluem:
  - Sistemas biométricos em tempo real que podem ser usados para qualquer tipo de vigilância, embora haja exceções para a prevenção de crimes e investigações criminais em contextos de aplicação da lei e segurança nacional.
  - Algoritmos de pontuação social que podem ser usados para avaliar indivíduos com base em características pessoais ou comportamento de uma forma que possa causar danos ou levar a um tratamento desfavorável de um indivíduo.
  - Sistemas manipuladores que exploram as vulnerabilidades de indivíduos específicos para distorcer seu comportamento de forma que possa causar danos físicos ou psicológicos.

- **Sistemas de AI de alto risco,** que estão listados no Anexo III e provavelmente constituirão a maioria dos sistemas de AI. Esses sistemas incluem:
  - Identificação biométrica e categorização de pessoas físicas [...].
  - Gestão e operação de infraestrutura crítica [...] [por exemplo, tráfego].
  - Educação e formação profissional [...].
  - Emprego e gestão de trabalhadores [...].
  - Acesso a serviços essenciais [...], incluindo a avaliação da capacidade de crédito, classificação de crédito ou priorização do acesso a esses serviços (Observação: isso se aplica especialmente aos sistemas de AI usados no setor de serviços financeiros).
  - Forças de segurança [...].
  - Gerenciamento de controles de fronteira [...].
  - Administração da justiça e processos democráticos [...].

- **Sistemas de AI de baixo risco [ou risco limitado],** que incluem sistemas que não usam dados pessoais ou fazem previsões que possam afetar direta ou indiretamente qualquer indivíduo, como aplicativos de manutenção preditiva industrial.

Com relação à interpretabilidade dos modelos de AI classificados como de alto risco, a Lei de AI estabelece[28] em seus artigos 13 e 14:

Art. 13. Transparência e comunicação de informações aos usuários

1. Os sistemas de AI de alto risco devem ser projetados e desenvolvidos de forma a garantir que operem com um nível suficiente de transparência para que suas informações de saída sejam corretamente interpretadas e utilizadas pelos usuários. [...]

2. Os sistemas de AI de alto risco devem ser acompanhados de instruções apropriadas para uso em formato digital ou outro formato apropriado, que devem incluir informações concisas, completas, corretas e claras que sejam relevantes, acessíveis e compreensíveis para os usuários. [...]

Art. 14. Vigilância humana

1. Os sistemas de AI de alto risco devem ser projetados e desenvolvidos de forma que possam ser efetivamente monitorados por pessoas físicas durante o período em que estiverem em uso, inclusive fornecendo-lhes uma ferramenta de interface homem-máquina adequada, entre outras coisas.

[...]

4. As medidas acima [...] devem permitir que as pessoas encarregadas da supervisão humana sejam capazes, dependendo das circunstâncias:
   a. **Compreender totalmente os recursos e as limitações do sistema de AI de alto risco** e controlar adequadamente seu funcionamento, de modo que possam detectar sinais de anomalias, mau funcionamento e comportamento inesperado e resolvê-los o mais rápido possível;
   b. estar ciente da possível tendência de confiar automaticamente ou em excesso nas informações de saída geradas por um sistema de AI de alto risco ("viés de automação") [...];
   c. interpretar corretamente as informações de saída do sistema de AI de alto risco [...];
   d. decidir, em uma determinada situação, não usar o sistema de AI de alto risco ou desconsiderar, invalidar ou reverter as informações de saída geradas por ele;
   e. intervir na operação do sistema de AI de alto risco ou interromper o sistema [...].

Como pode ser visto, o AI Act impõe condições restritivas sobre a interpretabilidade dos modelos de AI de alto risco (Fig. 4), que

[26]Espera-se que ela entre em vigor 20 dias após sua publicação no Diário Oficial da União Europeia e que seja de plena aplicação 24 meses após sua entrada em vigor.
[27]Floridi et al. (2022).
[28]Comissão Europeia (2021).

logo se tornarão obrigatórios em toda a União. Espera-se que isso desencadeie um número significativo de iniciativas de adaptação ao Regulamento, incluindo uma documentação mais abrangente dos modelos e seus usos, a aplicação de técnicas de interpretabilidade, o desenvolvimento de dashboards de monitoramento e de alerta de modelos ou a revisão do procedimento integrado para desenvolvimento, validação, implementação e uso de modelos, entre outros.

3. Diretrizes éticas para uma Inteligência Artificial confiável (Comissão Europeia)

Em abril de 2019, o Grupo de Especialistas de Alto Nível sobre AI da Comissão Europeia apresentou as Diretrizes Éticas para uma AI Confiável[29], após um processo de consulta com mais de 500 respostas do setor.

As Diretrizes propõem sete requisitos principais que os sistemas de AI devem atender para serem considerados confiáveis, que em resumo são: (i) ação humana e supervisão, (ii) solidez técnica e segurança, (iii) privacidade e gestão de dados, (iv) transparência, (v) diversidade, não discriminação e equidade, (vi) bem-estar ambiental e social e (vii) responsabilização.

Em particular, no que diz respeito à interpretabilidade dos modelos de AI, as Diretrizes declaram o seguinte como parte de seu requisito de transparência:

> 53. A explicabilidade é fundamental para conquistar e manter a confiança dos usuários nos sistemas de AI. Isso significa que os processos precisam ser transparentes, que os recursos e a finalidade dos sistemas de AI precisam ser comunicados abertamente e que **as decisões devem, na medida do possível, poder ser explicadas às partes que são** direta ou indiretamente **afetadas por elas.** Sem essas informações, não é possível contestar adequadamente uma decisão.
>
> **Nem sempre é possível explicar por que um modelo gerou um determinado resultado ou decisão** (ou qual combinação de fatores contribuiu para isso). Esses casos, que são chamados de algoritmos de "black box", exigem atenção especial.
>
> Em tais circunstâncias, **outras medidas relacionadas à explicabilidade** (por exemplo, rastreabilidade, auditabilidade e comunicação transparente sobre o desempenho do sistema) **podem ser necessárias, desde** que o sistema como um todo respeite os direitos fundamentais.
>
> **O grau de necessidade de explicabilidade depende, em grande parte, do contexto e da gravidade das consequências** de um resultado errôneo ou inadequado.

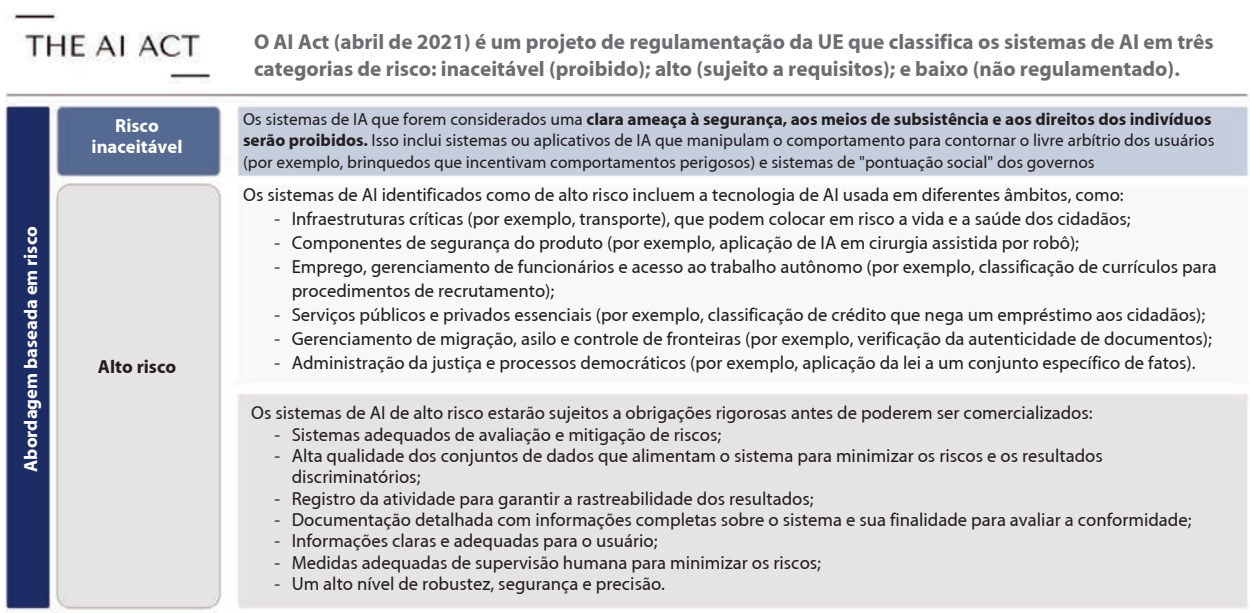
Como se pode ver, as Diretrizes apontam para a mesma direção: a exigência (que se eleva ao nível de necessidade ética) de que os modelos de AI sejam explicáveis.

Além disso, o que à primeira vista pode parecer um requisito mais relaxado para a interpretabilidade dos modelos de AI, uma vez que as Diretrizes reconhecem que alguns modelos de AI são mais difíceis de explicar, na verdade introduz uma complexidade adicional: a necessidade de classificar os modelos de AI de acordo com seu risco e seu potencial de interpretação, a fim de aplicar um grau maior ou menor de esforço em sua explicação.

Por fim, as Diretrizes têm como objetivo avaliar até que ponto um modelo de AI atende a esses sete requisitos, propondo uma lista de critérios de avaliação, que deve ser adaptada a cada caso específico. Com relação à explicabilidade, as Diretrizes formulam os seguintes critérios de avaliação[30], que devem ser integrados a outras ferramentas de avaliação já disponíveis para as organizações:

[29] Comissão Europeia (2019).
[30] Ibíd

Figura 4. Áreas de aplicação e requisitos do Artificial Intelligence Act.



- Você avaliou até que ponto as decisões e, portanto, o resultado produzido pelo sistema de AI são compreensíveis?
- Foi assegurado que é possível desenvolver uma explicação que seja compreensível para todos os usuários que desejam saber por que um sistema tomou uma decisão específica que levou a um resultado específico?
- Você avaliou até que ponto a decisão do sistema influencia os processos de tomada de decisão da organização?
- Você avaliou por que esse sistema específico foi implantado nessa área específica?
- Você avaliou o modelo de negócios do sistema (por exemplo, como ele cria valor para a organização)?
- Você desenhou o sistema de AI tendo em mente a interpretabilidade desde o início?
- Você pesquisou e tentou usar o modelo mais simples e mais interpretável possível para a aplicação em questão?
- Você já avaliou se pode analisar seus dados de treinamento e teste e se pode modificar e atualizar esses dados ao longo do tempo?
- Você avaliou se, após o treinamento e o desenvolvimento do modelo, tem alguma possibilidade de revisar sua interpretabilidade ou se tem acesso ao fluxo de trabalho interno do modelo.

4. Blueprint for an AI Bill of Rights (Casa Branca)

Em outubro de 2022, a Casa Branca propôs uma minuta da Declaração de Direitos sobre Inteligência Artificial[31], promovida pelo presidente Joe Biden e desenvolvida pelo Escritório de Política de Ciência e Tecnologia da Casa Branca (OSTP), e acompanhada de um manual (From Principles to Practice) sobre como implementá-la na prática.

O AI Bill of Rights estabelece cinco princípios ou direitos dos cidadãos em relação ao AI, que estão resumidos em[32]:

- Sistemas seguros e eficazes.
- Proteção contra discriminação de algoritmos.
- Privacidade de dados.
- Notificação e explicação.
- Processo alternativo de avaliação e correção humana em caso de falha de AI (fallback).

Em seu quarto princípio, referente à explicabilidade dos modelos de AI, ele afirma, entre outros, que[33]:

> Os projetistas, desenvolvedores e implementadores de sistemas automatizados devem fornecer documentação em linguagem simples e geralmente acessível, que inclua descrições claras da operação geral do sistema. [...]
>
> Os sistemas automatizados devem ser acompanhados de explicações que sejam tecnicamente válidas, significativas e úteis para você e para qualquer operador ou outras pessoas que precisem entender o sistema. [...]
>
> Os sistemas automatizados devem fornecer notificações de uso comprovadamente claras, oportunas, compreensíveis e acessíveis, além de **explicações sobre como e por que o sistema tomou uma decisão** ou executou uma ação.

5. Princípios sobre Inteligência Artificial (OECD)

Os Princípios da OCDE sobre Inteligência Artificial promovem o uso de AI que seja confiável e respeite os direitos humanos e os valores democráticos. Eles foram adotados em maio de 2019 pelos 38 países membros da OCDE. Foram os primeiros princípios desse tipo a serem endossados pelos governos e incluem recomendações concretas para políticas públicas e estratégias sobre AI.

Entre outros, eles afirmam que "os responsáveis da AI devem se comprometer com a transparência e a divulgação responsável dos sistemas de AI. Para esse fim, eles devem fornecer informações significativas, adequadas ao contexto e consistentes com o estado da técnica [...] para que aqueles afetados por um sistema de AI possam entender o resultado"[34]. O Observatório de Políticas de AI da OCDE, lançado em fevereiro de 2020, tem como objetivo ajudar os tomadores de decisão a implementar esses Princípios.

6. Discussion Paper on Machine Learning for IRB Models (EBA)

O Discussion Paper on Machine Learning for IRB Models, da Autoridade Bancária Europeia (EBA), publicado em novembro de 2021, é particularmente relevante para o setor bancário (Fig. 5).

O documento tem como objetivo analisar a relevância dos possíveis obstáculos à implementação de técnicas de machine learning no contexto da abordagem IRB para o cálculo de capital em instituições financeiras, inclui os desafios e os possíveis benefícios do uso dessas técnicas e estabelece determinados princípios e recomendações[35]. Um foco central do documento é, obviamente, como tornar o uso dessas técnicas compatível com a conformidade com o Regulamento de Capital Europeu (CRR[36]).

[31] OSTP da Casa Branca (2022).
[32] Ibíd.
[33] Ibíd.
[34] OECD (2019).
[35] Ver análise detalhada na Management Solutions (2021).
[36] CRR: Capital Requirements Rule (Regra de Requisitos de Capital), regulamentação central sobre capital em instituições financeiras na Europa.

Com relação à interpretabilidade dos modelos, o documento aborda essa questão sob o título "Concerns about the use of machine learning techniques" (Preocupações sobre o uso de técnicas de machine learning) e afirma[37]:

> As principais preocupações decorrentes da análise dos requisitos da CRR estão relacionadas à complexidade e à confiabilidade dos modelos de ML, em que os principais desafios parecem ser a interpretabilidade dos resultados, a governança, com referência específica ao aumento das necessidades de formação do pessoal, e a dificuldade de avaliar a capacidade de generalização de um modelo (ou seja, evitar o overfitting).
>
> Para entender as relações subjacentes entre as variáveis exploradas pelo modelo, os profissionais desenvolveram várias técnicas de interpretabilidade [...] [e] a escolha de qual dessas técnicas usar pode representar um desafio em si, pois essas técnicas geralmente permitem apenas uma compreensão limitada da lógica do modelo.

Além disso, o documento introduz a necessidade de se encontrar um equilíbrio entre a complexidade e a interpretabilidade do modelo e, diferentemente de outras regulamentações, desce a um nível mais técnico ao recomendar às instituições financeiras:

a. Analisar de forma estatística: i) a relação de cada variável de entrada com a variável de saída, ceteris paribus; ii) o peso global de cada variável de entrada na determinação da variável de saída, para detectar quais variáveis têm maior influência na predição do modelo. Estas análises são particularmente relevantes quando não é possível determinar uma representação próxima e pontual da relação entre a variável de saída do modelo e as variáveis de entrada devido à complexidade do modelo.

b. Avaliar a relação econômica de cada variável de entrada com a variável de saída para garantir que as estimativas do modelo sejam plausíveis e intuitivas.

c. Apresentar um documento de síntese que explique de forma simples o modelo baseado nos resultados das análises descritas na alínea a. O documento deve, no mínimo, descrever:

   i. Os principais fatores do modelo.
   ii. As principais relações entre as variáveis de entrada e as previsões do modelo.

   O documento é dirigido a todas as partes interessadas, incluindo a equipe que usa o modelo para fins internos.

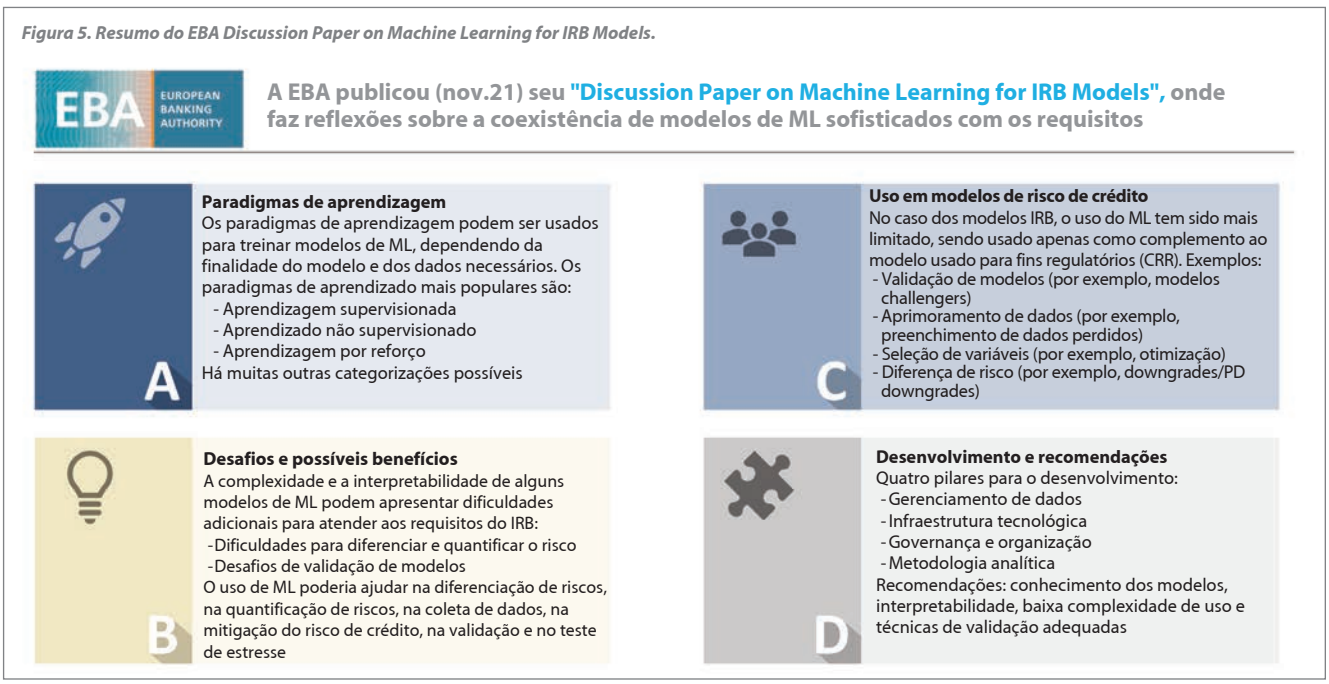
d. Garantir a detecção de possíveis vieses no modelo (por exemplo, overfitting à amostra de treinamento).

Na prática, enquanto o setor bancário aguarda a versão final do documento consultivo da EBA, a maioria das instituições que usam técnicas de machine learning em seus modelos IRB já está adaptando suas estruturas de desenvolvimento, monitoramento e validação de modelos para garantir a conformidade futura.

Um elemento comum em todas as referências regulatórias mencionadas acima, como pode ser visto, é a necessidade de fornecer uma explicação aos cidadãos sobre o uso da AI, e fazê-lo em dois níveis: a interpretabilidade e a transparência do modelo de AI como um todo, e a capacidade de explicar uma decisão específica do modelo, se necessário.

[37] EBA (2021).

Figura 5. Resumo do EBA Discussion Paper on Machine Learning for IRB Models.



Além das referências regulatórias descritas acima, há um grande número de publicações, princípios, diretrizes e projetos de regulamentação em várias jurisdições que abordam a interpretabilidade dos modelos de AI, tanto gerais quanto setoriais, e tanto regionais quanto locais em cada país; a seleção apresentada nesta seção inclui aqueles considerados como tendo o maior escopo e influência potencial.

Impactos na organização e nos processos

Um princípio essencial da XAI como disciplina é que, além do desenvolvimento de técnicas específicas de explicabilidade ou da construção de modelos inerentemente interpretáveis, essa explicabilidade e interpretabilidade devem ser integradas à organização e aos processos da empresa.

Colocado em prática, esse princípio implica o desenvolvimento e a implementação de um framework de XAI, que pode ser estruturado em quatro elementos:

1. Técnicas de interpretabilidade de modelos de AI
2. Integração aos processos de gestão de risco de modelo (MRM)
3. Suporte tecnológico
4. Fator humano

1. Técnicas de interpretabilidade dos modelos de AI

No centro de um framework de XAI estão as técnicas de interpretabilidade e explicabilidade, que podem ser resumidas em três aspectos:

- **Interpretabilidade do desenho do modelo:** isso inclui analisar como o modelo se comportaria em diferentes cenários (por exemplo, ataques adversários, cenários extremos...), entender como os submodelos e os conjuntos de modelos funcionam e integrar a interpretabilidade ao desenho do modelo aplicando restrições durante o desenvolvimento do modelo.

- **Interpretabilidade dos resultados do modelo:** refere-se à detecção de quais variáveis influenciam a previsão do modelo e como, por meio da interpretabilidade local (LIME, SHAP, etc.) e global (PDP, significância da variável, modelos substitutos, análise de sensibilidade); à avaliação do sentido econômico de cada variável (por exemplo, análise de caso de uso de uma amostra representativa de dados); e à garantia de que a documentação do modelo descreva corretamente o modelo, incluindo as variáveis de entrada e sua relação com os resultados.

- **Outros aspectos:** garantir a detecção de possíveis vieses no modelo (por exemplo, overfitting, dados de entrada tendenciosos, erros de dados) e monitorar regularmente o modelo, especialmente quando seu escopo mudar ou quando for aplicado a dados diferentes dos dados de desenvolvimento.

Devido à sua importância, as principais técnicas de interpretabilidade e explicabilidade serão desenvolvidas na seção a seguir.

2. Integração aos processos de gestão do risco de modelo (MRM)

A interpretabilidade dos modelos de AI é uma característica que transcende o desenvolvimento e afeta toda a cadeia do ciclo de vida do modelo e, portanto, todo o framework de gestão de risco do modelo. Um resumo não exaustivo da incorporação da XAI ao framework de MRM de uma empresa inclui a análise dos seguintes elementos:

- **Governança:** atualizar o framework de organização e de governança para incorporar a XAI; avaliar o impacto da regulamentação aplicável aos modelos de AI; atualizar o sistema de classificação de modelos para abordar a falta de interpretabilidade como um risco importante; atualizar o inventário de modelos e os procedimentos de inventário para incorporar elementos da XAI (por exemplo, atributos específicos para modelos de AI).

- **Desenvolvimento:** atualizar as políticas e os procedimentos de desenvolvimento de modelos, bem como os requisitos de documentação; avalie a imparcialidade e a parcialidade, a interpretabilidade das entradas, o design e as saídas, os dados, o risco de fornecedores, as métricas de capacidade preditiva, os limites para o uso de modelos de AI etc.; realizar a análise de sensibilidade dos modelos de AI para identificar vulnerabilidades; incluir no framework de desenvolvimento testes específicos para XAI.

- **Monitoramento:** atualizar o framework de monitoramento de modelos e completá-la com testes específicos de XAI; revisar limites e ações para não conformidade; desenvolver sistemas de alerta antecipado para detectar mudanças nos modelos de AI; revisar a conformidade com o apetite de risco de modelo; avaliar a necessidade de desenvolver um

módulo de monitoramento ad hoc para modelos de aprendizado dinâmico (ou seja, modelos que se recalibram automaticamente sem intervenção humana).

- **Validação:** atualizar o framework de validação interna para detectar possíveis riscos associados aos modelos de AI e incorporar testes de XAI; estabelecer um framework de validação cruzada para garantir a qualidade dos modelos de AI; avaliar o impacto das mudanças no ambiente de produção nos modelos de AI.

- **Implementação:** atualizar o processo de implementação do modelo para incorporar testes específicos às características da XAI; atualizar, quando apropriado, a plataforma tecnológica para permitir a produção de modelos de AI.

- **Uso:** atualizar procedimentos para o uso de modelos de AI para determinar sua adequação ao contexto em que serão usados; revisar e concluir o treinamento de usuários em modelos de AI; atualizar protocolos para detectar possíveis situações de uso indevido ou exploração de modelos.

- **Auditoria:** implementar um framework de auditoria para modelos de AI para garantir sua implementação e uso adequados; estabelecer testes de XAI para a auditoria de modelos de AI; avaliar a adequação dos sistemas de controle interno para garantir a qualidade dos modelos de AI; analisar trilhas de auditoria para detectar possíveis riscos associados aos modelos de AI.

Portanto, o uso de modelos de AI implica uma revisão completa das políticas e dos procedimentos durante todo o ciclo de vida do modelo para incorporar, no mínimo, os elementos da XAI.

3. Suporte tecnológico

A implementação de um framework de XAI tende a começar com ferramentas departamentais e, assim que atinge um nível mínimo de maturidade, requer soluções tecnológicas profissionais para dar suporte aos aspectos de interpretabilidade dos modelos de AI.

Essas soluções podem ser classificadas em dois grupos:

- Interpretabilidade: desenvolvimento de sistemas que implementem técnicas de interpretabilidade de forma padronizada e homogênea. Eles devem permitir que a interpretação dos modelos seja realizada automaticamente, facilmente configurável e com alta qualidade, incorporando as técnicas mais comuns e oferecendo flexibilidade para adicionar novas técnicas à medida que forem desenvolvidas[38].

- Governança de modelos: desenvolvimento ou atualização de sistemas de governança de modelos para dar suporte aos aspectos de XAI em MRM (inventário, classificação, documentação etc.), garantindo assim que os modelos disponíveis atendam aos requisitos de qualidade, segurança e explicabilidade exigidos[39].

Além disso, recomenda-se uma abordagem holística que englobe todos os aspectos do framework de XAI. Isso inclui o uso de ferramentas de análise de dados, o desenvolvimento de APIs para a integração dos sistemas de interpretabilidade e governança de modelos descritos acima, a criação de mecanismos de segurança e auditoria e a definição de protocolos para garantir a conformidade com os padrões de qualidade e explicabilidade.

4. Fator humano

Um quarto elemento na integração da XAI à organização e aos processos é a consideração do fator humano. Isso inclui, entre outros:

- **Recrutamento e retenção de talentos:** desenvolvimento de programas de recrutamento e retenção de talentos especializados em XAI, para garantir a presença de profissionais com o conhecimento técnico e a experiência necessários para aplicar XAI na empresa, o que é especialmente relevante em um mercado de trabalho com escassez desse perfil profissional.

- **Treinamento:** desenvolvimento de programas de treinamento para equipes de desenvolvimento de modelos de AI, equipes de governança de modelos e usuários de modelos de AI para garantir que todos os envolvidos entendam os princípios básicos da XAI e como aplicá-los no contexto específico da empresa.

- **Cultura:** desenvolver uma cultura empresarial que incentive o uso e a exploração da explicabilidade e interpretabilidade dos modelos de AI. Isso pode incluir a adoção de metodologias ágeis para o desenvolvimento de modelos de AI, a criação de uma cultura de colaboração entre as equipes de desenvolvimento e governança de modelos e a consideração da explicabilidade como um fator crítico na aprovação de modelos de AI.

- **Gestão de mudanças:** desenvolvimento de programas de gestão de mudanças para garantir a adoção adequada da XAI pelas equipes da empresa que trabalham com modelos de AI. Isso inclui a motivação das equipes de desenvolvimento, a análise dos custos e benefícios da explicabilidade, a definição de protocolos de comunicação com terceiros, etc.

Em conclusão, a explicabilidade e a interpretabilidade dos modelos de AI são aspectos fundamentais que precisam ser integrados à organização e aos processos da empresa por meio de um framework de XAI adequado e abrangente, o que é essencial para garantir o uso desses modelos de acordo com a regulamentação e as boas práticas.

[38] Nesse sentido, a Management Solutions tem o ModelCraft™, um sistema proprietário de modelagem de componentes e AutoML, que incorpora um módulo completo de interpretabilidade. Ver Management Solutions (2023).

[39] A Management Solutions também possui o Gamma™, um sistema proprietário de governança de modelos que abrange todos os aspectos acima. Ver Management Solutions (2022).