PRADO VIDIGAL

Privacidade & Digital

Transparência e percepção de justiça em sistemas de inteligência artificial: **como adotar explicações eficientes?**

Sistemas de **inteligência artificial (IA)** estão cada vez mais presentes em nosso dia a dia, por exemplo, em processos seletivos, avaliações de crédito, chatbots de atendimento etc. Essa popularização também chama atenção para **debates sobre a relação entre ganho de eficiência** e **impactos do uso de tais sistemas**.

Nesse contexto, a **transparência algorítmica** assume papel fundamental, uma vez que a capacidade de explicar o raciocínio e os resultados de um sistema ("explicabilidade") pode **impulsionar a construção de confiança** e aumentar a percepção de justiça por parte de usuários, **impactando sua vontade** de utilizar ou não o sistema.

Assim, sem pretensão de exaurir o tema, elaboramos este material para **auxiliar organizações na adoção de práticas que impulsionam a transparência e a confiança em atividades envolvendo sistemas de IA.**

1 Como desenvolver explicações sobre o funcionamento de sistemas IA?

É importante que a percepção dos usuários seja levada em consideração no momento da escolha do método de explicação mais apropriado para um sistema; determinados estilos podem se revelar mais eficazes para alguns sistemas e/ou *stakeholders* específicos.

Também é necessário considerar que, atualmente, os usuários são bombardeados com diversas informações sobre diferentes aspectos de produtos/serviços e, consequentemente, acabam não se interessando por leituras extensas e excessivamente técnicas. Por isso, organizações devem pensar em diferentes estratégias de comunicação, por exemplo, utilizando conceitos familiares para redução do esforço cognitivo dos usuários[1].

Além disso, é importante que organizações estejam atentas a limitações para o fornecimento de informações referentes a metodologia, critérios e bancos de dados utilizados, evitando a divulgação de segredo comercial e industrial. Inclusive, o fornecimento ilimitado de informações acerca do funcionamento de determinado sistema de IA pode desencadear violações à segurança dos próprios usuários, pela não preservação da confidencialidade de detalhes que podem colocar em risco a segurança e robustez dos sistemas.

Assim, abordamos abaixo algumas reconhecidas técnicas para apresentar explicações, que, inclusive, podem ser utilizadas de forma combinada[2]. Os exemplos práticos foram extraídos – de forma adaptada – de experimento no qual os participantes foram submetidos a um sistema de análise de candidaturas e recomendação de contratação para vaga de emprego.

1) Explicação baseada em casos:

apresenta-se um caso sobre o funcionamento/aplicação do sistema de IA em situação semelhante a do usuário.

Exemplo prático:

"Um caso semelhante, que recebeu o mesmo resultado, é de um candidato que era um estudante com desempenho médio e com alguma experiência relevante para o trabalho. Ele foi positivamente recomendado por colegas. Como o candidato tinha currículo e personalidade semelhantes à sua, os resultados dos testes foram similares."

[1] GEDIKLI, Fatih; JANNACH, Dietmar; GE, Mouzhi. How should I explain? A comparison of different explanation types for recommender systems. International Journal Of Human-Computer Studies, [S.L.], v. 72, n. 4, p. 367-382, abr. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.ijhcs.2013.12.007. Disponível em: https://doi.org/10.1016/j.ijhcs.2013.12.007. Acesso em: 18 abr. 2023.

[2] As técnicas e os exemplos apresentados – de forma adaptada – foram extraídos do estudo "Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. SHULNER-TAL, Avital; KUFLIK, Tsvi; KLIGER, Doron. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users perceptions of fairness toward an algorithmic system. Ethics And Information Technology, [S.L.], v. 24, n. 1, p. 1-13, 24 jan. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s10676-022-09623-4. Disponível em: https://dl.acm.org/doi/abs/10.1007/s10676-022-09623-4. Acesso em: 18 abr. 2023.

2) Explicação com base demográfica:

apresenta-se dados demográficos agregados (como idade, sexo, nível de rendimentos ou ocupação) que influenciam no treinamento do sistema de IA e, consequentemente, em seus resultados.

Exemplo prático:

"Os resultados demonstram que 17% dos candidatos que estão classificados no top 10% de suas turmas de graduação são positivamente recomendados pelo sistema o sistema e 36% dos candidatos com 10 anos de experiência relevante são negativamente recomendados pelo sistema."

3) Explicação baseada na influência dos inputs:

apresenta-se quais inputs/características são decisivas para obtenção dos resultados.

Exemplo prático:

"Nosso sistema avalia a possibilidade de o candidato progredir no processo seletivo. Os fatores relevantes podem afetar os resultados positiva (+) ou negativamente (-), conforme abaixo:

- Classificação da universidade (++).
- Classificação do candidato na universidade (+).
- CV do candidato (+).
- Resultados do teste de personalidade do candidato (-).
- Cartas de recomendação do candidato (--)."

4) Explicação baseada na sensibilidade:

apresenta-se uma análise que explica como diferentes alterações nos inputs podem modificar o resultado final gerado pelo sistema.

Exemplo prático:

"Nosso sistema pode apresentar resultados diferentes a depender das alterações de inputs, conforme abaixo:

- Se o candidato tivesse mais um ano de experiência relevante para este trabalho, a probabilidade de uma recomendação positiva pelo sistema seria aumentada em 34%.
- Se o candidato fosse classificado entre os top 10% de sua turma de graduação, a probabilidade de uma recomendação positiva pelo sistema seria aumentada em 23%."

5) Explicação baseada em certificação:

apresenta-se os resultados de um processo de auditoria do sistema de IA.

Exemplo prático:

"Nosso sistema foi testado em processo de auditoria para verificar a equidade de resultados em relação à diferentes segmentos da população, tendo sido verificado que os requerimentos exigidos foram implementados."

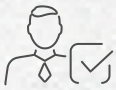
2 Como avaliar a efetividade das explicações?

Após a construção da explicação, é importante avaliar sua eficácia. Para tanto, avanços relevantes têm sido feitos no âmbito do campo de estudo denominado *Explainable Artificial Intelligence* (XAI). Por exemplo, o XAI Test[3] é uma metodologia de avaliação da utilidade de explicações a partir da percepção de usuários, buscando avaliar os parâmetros a seguir:

- A explicação abrange toda a informação relevante para se tomar uma decisão?
- A explicação ajuda a tomar uma decisão de forma mais rápida?
- A explicação é útil para se tomar uma decisão?

Além disso, outros parâmetros podem ser utilizados para avaliar a efetividade de explicações, por exemplo[4]:

A explicação apresenta detalhes e granularidade suficiente para compreensão do funcionamento do sistema de IA.



Usuários não precisam de suporte/assistência para compreender as informações apresentadas.

As explicações não apresentam inconsistências.

As explicações não necessitam de referências extras para serem compreendidas, como regulações, guias externos etc.

As explicações são fornecidas em tempo razoável.

3 Quais formas de explicação são mais eficientes para aumentar a percepção de justiça?

De acordo com estudos realizados a partir das técnicas de explicação aplicadas ao experimento mencionado anteriormente, **o resultado dos sistemas de IA é determinante para a percepção de justiça por parte dos usuários:**

Em casos de recomendação negativa, isto é, quando o usuário não foi recomendado para a vaga de emprego pelo sistema, este foi considerado injusto independentemente da explicação. Ainda assim, os usuários apontaram que as explicações ajudaram a compreender o resultado do sistema.

A técnica de explicação baseada na sensibilidade foi julgada pelos participantes como a melhor, tanto nos aspectos de percepção de justiça, como de compreensão do funcionamento do sistema.

Em casos de recomendação positiva, isto é, quando o usuário foi recomendado para a vaga de emprego pelo sistema, este foi considerado justo, exceto quando a explicação com base demográfica foi utilizada. Os usuários também pontuaram que as explicações auxiliaram na compreensão do funcionamento do sistema.

A técnica de explicação baseada em certificação foi a mais bem avaliada em termos de percepção de justiça, pois, ao trazer certa imparcialidade para a análise, auxilia na construção de confiança junto ao usuário.

• • •

[3] JESUS, Sérgio; BELÉM, Catarina; BALAYAN, Vladimir; BENTO, João; SALEIRO, Pedro; BIZARRO, Pedro; GAMA, João. How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. Disponível em: https://arxiv.org/abs/2101.08758. Acesso em 18 abrl. 2023.

[4] HOLZINGER, Andreas; CARRINGTON, André; MULLER, Heimo. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. Springer/Nature KI Kuenstliche Intelligenz 34, 193-198 (2020). Disponível em: https://arxiv.org/abs/1912.09024. Acesso em 18 abrl. 2023.

Portanto, é de se concluir que o resultado gerado pelo sistema de IA (positivo ou negativo) afeta significativamente as percepções de justiça por parte dos usuários. Ainda assim, nota-se que o fornecimento de explicações auxilia na compreensão do funcionamento dos sistemas (tanto em resultados "positivos", como em resultados "negativos"), reforçando a transparência.

Considerando-se que o resultado do sistema é determinante para a percepção de justiça, é importante garantir, desde o momento de concepção, que o sistema de IA gere resultados que possam ser defendidos como justos. Nesse contexto, o desenho de uma estrutura de governança em IA é essencial para promoção de uma gestão eficaz de riscos e desenvolvimento de sistemas mais confiáveis.

Como forma de introduzir rotinas de *accountability*, organizações podem incorporar práticas como:

Elaboração de **avaliações de impacto em sistemas de inteligência artificial** para identificação de possíveis consequências de uma determinada iniciativa sobre interesses socialmente relevantes.

Estruturação de **comitês de ética**, que funcionam como mecanismos de supervisão no desenvolvimento e uso de sistemas de IA e podem contar com formação diversificada para ampliação de visões e perspectivas de justiça e equidade.

Realização de **auditorias regulares** (internas e externas), que podem ser focadas em aspectos como transparência, ética e vieses discriminatórios.

Evidentemente, este é um tema em constante desenvolvimento e não há uma solução única que possa ser aplicada para toda e qualquer organização, sendo essencial acompanhar sua evolução. A definição de estratégias para promoção de transparência e percepções de justiça em relação ao uso de sistemas de IA deve levar em consideração as particularidades de cada organização, especialmente modelo de negócios, recursos, cultura e apetite a riscos.

De todo modo, nos parece válido conhecer e explorar as técnicas que têm sido ventiladas para fins de explicação e os resultados, ainda que preliminares, de sua eficácia, como ponto de partida para a escolha e desenvolvimento do método mais adequado para determinado sistema em análise.

Para maiores informações sobre o tema, entre em contato com nosso time.

*Material produzido por **Prado Vidigal Advogados** em maio de 2023.*

Autores(as):
Carolina Giovanini
Paulo Vidigal