Responsabilidade Distribuída, Casos Concretos, Backlogs Éticos, a Origem do Risco e a Inteligibilidade das Decisões

Uma Aula Reprodutível

Robert Junqueira
martinsjunqueira@direito.uminho.pt
Centro de Investigação em Justiça e Governação
da Escola de Direito da Universidade do Minho

Desenho diretriz de aula, preparado no seguimento de uma interpelação da Professora Doutora Irene Portela, para a Unidade Curricular de Ética da Inteligência Artificial do Mestrado de Inteligência Artificial Aplicada da Escola Superior de Tecnologia da Universidade Politécnica do Cávado e do Ave.

Originalmente ministrada em língua inglesa por Robert Junqueira
no contexto da referida UC, a 20 de maio de 2026, das 18h00 às 20h00, em Barcelos.

http://doi.org/10.6084/m9.figshare.32347194

Quem falha quando a IA falha?

Vamos começar com uma situação empresarial, não com uma definição. Imaginem que trabalham numa organização que está prestes a lançar uma funcionalidade com IA. A promessa é boa: poupar tempo, reduzir custos e tomar decisões mais consistentes.

Pode ser uma IA para ordenar candidatos a emprego, recomendar crédito, responder automaticamente a reclamações ou sinalizar clientes com maior risco de fraude.

A equipa está entusiasmada. A direção quer lançar depressa, porque um concorrente já anunciou algo parecido. O fornecedor garante que o modelo é robusto. A equipa de produto diz que os testes estão aceitáveis. O marketing já tem uma campanha pronta.

Isto soa-vos plausível numa empresa real? Imagino que sim.

Agora imaginem que, duas semanas depois do lançamento, alguém é prejudicado. Um candidato nunca chegou à entrevista. Um cliente foi recusado sem perceber porquê. Uma reclamação urgente ficou enterrada. Uma pessoa descobre que foi tratada como "de risco" por uma classificação que ninguém sabe explicar bem.

Quem é responsável? Quem é responsável se alguma coisa correr mal?

Quem é responsável se alguma coisa correr mal? Quem é responsável?

Vamos ver.

90 segundos. Produzam uma breve lista. Escrevam tudo o que vos vier à cabeça a nível dos atores, sistemas e regras envolvidos na decisão da empresa antes do resultado aparecer: indivíduos, grupos, dados, objetivos de negócio, regras internas, métricas, prazos...

Quem tocou na decisão antes do *output* aparecer? Quem comprou ou aprovou o sistema? Quem escolheu e curou os dados? Quem definiu os objetivos de negócio? Quem usa o *output* no dia a dia? Quem sofre a decisão sem participar nela?

90 segundos.

Pessoas/equipas	**Sistemas/dados**	**Regras/pressões**	**Ausências**
Direção	Modelo	Prazo de lançamento	Pessoas afetadas
Responsável pelo produto	Dados históricos	KPI/redução de custo	Auditores externos
Cientistas de dados	Fornecedores	Política Interna	Reguladores
Jurídico/compliance	Dashboard	Contrato com fornecedor	Canal de reclamação
Outros funcionários	*Logs*	Campanha de marketing	Representantes de utilizadores

A nossa lista já ultrapassa bastante a palavra "algoritmo".

Isto já não é só um modelo. É uma cadeia de decisão. Há muitos responsáveis, mas responsabilidade distribuída não é responsabilidade diluída.

A IA pode ser a parte mais visível, mas antes dela houve dados escolhidos, objetivos definidos, fornecedores contratados, métricas aceites, testes considerados suficientes e uma decisão de lançar o produto. Depois há as pessoas afetadas ou potencialmente afetadas, explicações a dar, revisões a serem feitas e responsabilidades a assumir.

A IA falha — ou funciona bem — no quadro de uma cadeia de decisão.

Exemplo rápido: uma empresa recebe 3000 candidaturas por mês e compra uma IA para ordenar candidatos por "probabilidade de sucesso". O fornecedor treinou o sistema com dados históricos da própria empresa. A equipa de RH diz que é só uma ajuda; na prática, porém, entrevista-se quase sempre quem aparece no topo da lista. A direção quer reduzir o tempo de contratação.

Se candidatos com percursos não lineares, universidades menos conhecidas, pausas profissionais, pessoas do género feminino em idade reprodutiva, portugueses, franceses, chineses, brancos, negros ou amarelos começam a ficar sistematicamente para trás, dizer "foi o modelo" é fugir à responsabilidade.

A decisão veio de uma cadeia: dados históricos, objetivo de rapidez, ranking, uso quase automático por parte da equipa de RH, entre outros fatores.

Então, hoje, "Ética da IA" não significa "ter uma opinião bonita (ou politicamente correta) sobre tecnologia".

Ética da IA é uma tarefa clara e metódica: organizar a responsabilidade dentro de uma cadeia de decisão; repito: a tarefa central da ética da IA consiste em organizar a responsabilidade dentro de uma cadeia de decisão, seja preventivamente, antes de haver qualquer dano, seja para atribuir responsabilidade pelo dano, ou seja mesmo como mero exercício de compreensão da nossa reponsabilidade a da dos outros.

A ética é uma dimensão transversal das nossas vidas pessoais e profissionais que pode e deve ser assiduamente convocada.

Atividade 1

Vamos começar com a primeira atividade individual. Cada um vai construir um caso concreto. O que é um caso? Não é um tema geral como "viés na IA"; é uma situação real: uma empresa, uma decisão, pessoas afetadas, e por aí em diante. Apontem os passos.

Passo 1

Escolher um tema entre estes quatro: viés, privacidade, explicabilidade e automação.

Passo 2

Imaginar uma empresa e uma decisão concreta que a IA vai ajudar a tomar. Por exemplo: uma IA que ordena candidatos a emprego antes da entrevista; ou uma IA que recomenda aprovar ou recusar um pedido de crédito; ou uma IA que decide que reclamações de clientes passam para um humano; ou uma IA que prevê quais clientes vão cancelar o serviço. Discriminem qual é a decisão em causa e quem é que usa essa decisão na empresa (pode ser a equipa de RH, a direção, enfim...).

Passo 3

Perguntem: quem é abrangido por esta decisão? Atenção que nem sempre é só quem usa a ferramenta.

Passo 4

Qual é o benefício almejado? A empresa quer o quê? Reduzir tempo? Cortar custos? Ser mais consistente?

Passo 5

Qual é um dano plausível? Isto é: o que pode correr mal que prejudica alguém?

Depois destas cinco respostas, vão escrever uma frase chamada *user story ética*. Atenção: o *user* aqui é interno à empresa, é quem "assina" com base em processos da IA. É um modelo útil que força a ser concreto. Usem este template:

'Enquanto [posição – i.e., quem quer usar a IA], queremos usar IA para [decisão], com vista a [benefício], desde que [condição ética mínima].'

A condição ética mínima é um cinto de segurança. Não pode ser uma palavra bonita como "justiça" ou "transparência". Tem de ser algo verificável. Exemplo: *desde que nenhum candidato seja excluído sem possibilidade de revisão humana da decisão;* ou *desde que o cliente receba uma explicação simples e um contacto para reclamar*.

Temos 5 minutos para fazer isto. Antes de começarmos, tomem nota do seguinte:

Tema vs caso	Tema é conversa de café ("viés é mau"). Caso é decisão concreta ("esta IA ordena candidatos e pode excluir grupos"). Sem caso, não há trabalho.
Decisão assistida vs automatizada	Assistida: humano decide com ajuda da IA. Automatizada: IA decide e humano só vê se houver reclamação. Isso muda a consideração do risco.
Dano plausível	O conteúdo do risco. Não é necessariamente uma catástrofe, mas é algo inaceitável que pode acontecer, por exemplo, logo na primeira semana.
Condição ética mínima	É um cinto de segurança. Não resolve tudo, mas obriga a equipa a dizer "não lançamos se isto falhar". Se a condição for vaga, não traz segurança alguma.

Agora vamos ouvir cada um. Dois minutos por pessoa. Digam:

- qual é o vosso caso (empresa + decisão)
- quem é afetado
- o benefício almejado
- o dano plausível
- e a condição ética mínima

Agora vamos pegar na condição ética mínima de cada um e transformá-la em trabalho.

Atividade 2

Vamos fazer o *backlog* ético. O que é um *backlog*? É uma lista de trabalho *para ontem*. Em ética, quando somos chamados a agir é porque já estamos atrasados. Toda a tarefa ética entra automaticamente no *backlog*. O *backlog* não é uma lista de boas intenções. É algo do tipo: temos de fazer X, testar Y da maneira W pela razão Z, documentar o processo e/ou tomar decisões, e.g. antes de lançar o produto.

A diferença entre trabalho e belos sentimentos é esta: "ser justo" não é trabalho; "testar o impacto por grupo de candidatos" já é trabalho. "Ser transparente" não é trabalho; trabalho é "escrever uma explicação sincera e em linguagem acessível para o cliente e criar uma estratégia de revisão (ou um canal de reclamações)".

Cada item de um *backlog* ético precisa de quatro coisas:

- **Risco**: o que pode correr mal. Já sabem o risco ou algum risco do vosso caso concreto.
- **Evidência necessária**: o que prova que o risco está controlado. Não é opinião, é algo que alguém pode verificar.
- **Responsável**: uma pessoa ou função responsável pelo item de *backlog*. Se ninguém é responsável, ninguém faz.
- **Critério de aceitação**: como sabemos que há conformidade. Tem de ser verificável.

Vão desenhar, no mínimo, dois itens de *backlog ético* para o vosso caso concreto. Além disso, vão identificar pelo menos um item como "bloqueador". Um bloqueador é algo inegociável, que se não for satisfeito não se pode sequer pensar em lançar sequer uma versão piloto. Se falharmos relativamente ao item bloqueador, o projeto não avança. Isto é importante porque, numa empresa, a pressão para lançar é grande. Os bloqueadores são os nossos guardiões máximos do templo da ética.

Um exemplo. Imaginem o caso dos candidatos. Um item de *backlog* poderia ser:

> Há risco de exclusão sistemática de grupos, pelo que o produto não é eficaz sem revisão humana intensiva e, portanto, não traz benefícios. Faz-se necessária evidência que afaste o risco: e.g. um relatório que mostra a posição média de cada grupo no ranking. O responsável por isto é o data lead (ou a equipa de data science). O critério de aceitação ética do produto é nenhum grupo ter posição média significativamente pior que os outros antes da versão piloto.

No fim de cada item, vem a pergunta: este item é um bloqueador ético? Se, no caso do exemplo, a resposta for sim, a conclusão operacional é que "Se houver diferença injustificada entre grupos, não lançaremos o produto".

Antes do exercício, tomem nota de cada uma destas definições:

Agora, cada um de vocês vai pegar no caso e na condição ética que escreveram há pouco e vai criar entre 2 e 4 itens de *backlog*, pelo menos um deles sendo um bloqueador. Em cada item, devem dizer qual é o risco, a evidência necessária para considerarmos a situação controlada, quem é responsável pelo item, qual o critério de aceitação que a evidência necessária implica e, ainda, se o item é bloqueador.

Temos 10 minutos para fazer isto.

Backlog ético	Lista de tarefas que tornam a condição ética operacional. Sem backlog, a ética fica no discurso.
Evidência	Algo verificável por uma segunda ou terceira pessoa. Pode ser um relatório, um log, um teste, um procedimento documentado. Não é "confiamos no fornecedor".
Responsável	Não é culpado se correr mal. É a pessoa que se compromete a fazer o item avançar. Pode ser uma função (ex: "compliance"), mas o ideal é uma pessoa.
Critério de aceitação	Condição binária: ou está cumprido ou não. Ex: "relatório entregue e aprovado" é critério. "Mais ou melhor justiça" não é.
Bloqueador	Item que, se falhar, impede o lançamento. Isto dá força à ética: não é um conselho, é uma condição contratual interna.

Agora, cada um vai falar 3 minutos. Digam: item, risco, evidência necessária, responsável, critério de aceitação, se é bloqueador e a razão para ser ou não ser.

Bom, já temos casos e *backlogs*. Agora vamos ver onde esses riscos nascem dentro do sistema de trabalho. Vamos desenhar um mapa de atividade e, depois, pensar na inteligibilidade das decisões.

Atividade 3

Vamos à última atividade prática. Ela junta duas coisas que ainda não fizemos: localizar a fonte do risco no sistema e pensar na inteligibilidade da decisão para quem precisa de explicações.

Primeiro, cada um vai desenhar um mapa de atividade. Um mapa de atividade é uma forma de não ficarmos cegos pela luz do algoritmo. Muitas vezes, achamos que o problema está no modelo. Mas o modelo é só uma peça. O risco pode vir dos dados, da regra de negócio, da pressão de prazo, da equipa que usa a recomendação de forma automática ou da desconsideração, ao longo do processo, da pessoa afetada.

Apontem a resposta a cada uma destas perguntas com base no vosso caso concreto. A maioria das respostas já estão nas vossas mãos:

- **Quem age?** Quem é a equipa ou pessoa que usa a IA?
- **Qual é a decisão?** O que é que a IA ajuda a decidir?
- **Com que instrumentos?** Dados, modelo, fornecedor, painel de controlo?
- **Sob que regras ou pressões?** Prazos, KPIs, leis, políticas internas, contratos?
- **Quem é afetado?** Há alguém, mesmo que não participe na decisão, que será afetado pelas decisões (ex: candidato, cliente, trabalhador)?

Cinco minutos para escreverem a resposta a estas perguntas, mais a pergunta: em que região nasce(m) o(s) risco(s) que identifiquei?

Estas perguntas e as respetivas respostas são as regiões do vosso mapa de atividade. Agora vamos localizar a origem do risco. Onde nasce o(s) risco(s) que identificaram? Por favor respondam, um de cada vez:

O risco de [X] nasce em [região do mapa] e exige responsabilidade de [alguém].

Muito bem. Não esquecer: o risco pode nascer nos dados históricos, nos KPIs de velocidade, na falta de revisão humana, no contrato com fornecedor, na ausência da pessoa afetada. O risco, como a culpa, não é diretamente atribuível ao algoritmo.

Agora vamos à inteligibilidade da decisão.

Vão escolher uma pessoa ou posição que precisa de ser capaz de compreender o processo de decisão. Pode ser a pessoa afetada (ex: candidato, cliente), pode ser uma questão de *compliance*, pode ser a equipa que usa a IA, pode ser a direção. Cada destinatário precisa de uma explicação diferente, talhada para o seu perfil. Não se trata de despejar informação técnica, que só é inteligível para especialistas: trata-se de dar algo que a pessoa percebe e que lhe permite agir. Escrevam:

Para [destinatário], a explicação suficiente é [o quê], em [formato: mensagem, relatório, dashboard, reunião], para que ele/ela possa [ação: contestar, auditar, corrigir, decidir].

Um exemplo: para um candidato rejeitado, a explicação é "a IA ajudou a ordenar candidaturas, o motivo principal foi a falta de experiência na área. Pode pedir revisão humana da decisão por via deste contacto".

Escolham um caso diferente do do exemplo e escrevam. Têm 5 minutos.

Por favor apresentem o que escreveram.

Muito obrigado. Agora mantenham presente que a inteligibilidade da decisão não vem do detalhamento técnico: vem do facto de haver uma explicação tal que permite à inteligência que capta a decisão por via da explicação ser capaz de agir de forma adequada aos seus potenciais propósitos com base na explicação dada. Se é, num dado caso específico, então estamos perante um processo de decisão inteligível no que respeita ao caso em apreço.