OTÁVIO MORATO
DIERLE NUNES

INTELIGÊNCIA ARTIFICIAL:
o desafio da explicabilidade

2025

EDITORA JusPODIVM
www.editorajuspodivm.com.br

1

BREVE HISTÓRIA DA IA

Em 18 de junho de 1956, um seminário realizado na Faculdade de Dartmouth (Nova Hampshire, EUA) começou com uma proposta ousada: investigar como as máquinas poderiam emular a inteligência humana. O evento, chamado de Conferência sobre Inteligência Artificial (IA) era liderado por John McCarthy (de Dartmouth), Marvin Minsky (Harvard), Nathaniel Rochester (IBM) e Claude Shannon (Bell Labs), todos pesquisadores norte-americanos de tecnologia e entusiastas da computação.

Os organizadores estavam confiantes de que, no futuro, seria possível desvendar a lógica, o raciocínio e o aprendizado, possibilitando a *reprodução artificial da inteligência humana*. O convite para o workshop refletia esse otimismo,

assumindo que "todo aspecto do aprendizado ou qualquer outra característica da inteligência pode, a princípio, ser tão precisamente descrito que uma máquina seria capaz de simulá-lo".

Figura 1: Um raro registro com membros da Conferência de Dartmouth sobre IA de 1956. Da esquerda para a direita: Oliver Selfridge, Nathaniel Rochester, Ray Solomonoff, Marvin Minsky, Trenchard More, John McCarthy e Claude Shannon. Fonte: IEEE Spectrum.

Durante cerca de um mês, o pequeno grupo de 10 debatedores mergulhou em discussões interdisciplinares, que misturavam matemática, filosofia, neurociência e engenharia. Foi ali que John McCarthy e seus colegas fizeram o primeiro uso científico do termo "inteligência artificial". A escolha da nomenclatura era, antes de tudo, uma declaração de propósito: era preciso inaugurar um ramo independente de estudos, com o objetivo de desenvolver *máquinas pensantes*. Embora não tenha resultado em descobertas práticas imediatas, a Conferência de Dartmouth foi um manifesto de possibilidades, marcando o nascimento oficial da IA e lançando alicerces teóricos para décadas de pesquisa.

Podemos definir inteligência artificial como um programa ou tecnologia criada para funcionar sozinha ou com

pouca interferência humana[1]. Ela aprende a partir das informações que recebe e pode se ajustar com o tempo para melhorar seu desempenho. Seu objetivo é analisar dados e gerar respostas úteis, como previsões, sugestões, criação de conteúdos ou decisões que podem afetar tanto o mundo digital quanto o físico.

INTELIGÊNCIA ARTIFICIAL

sistema **computacional** projetado para operar com diferentes níveis de **autonomia** e capaz de apresentar certo grau de **adaptabilidade** após sua implementação.

Esses sistemas processam dados recebidos e, a partir deles, são capazes de gerar saídas como previsões, recomendações, criação de conteúdos ou decisões que influenciam ambientes físicos ou virtuais.

Fundamentada em *algoritmos* – conjuntos de regras e instruções que permitem processar informações e resolver problemas – a IA vai além da execução mecânica, empregando dados para solucionar questões complexas e imitar características humanas, como percepção, raciocínio e adaptação ao meio. Trata-se, portanto, da convergência entre *tecnologia computacional* e *aprendizado*; por meio do segundo, a primeira amplia a capacidade das máquinas de operar de maneira estratégica e responsiva.

O desafio de traduzir o pensamento humano em sistemas lógicos é antigo, como vemos na lógica Aristotélica e

1. Utiliza-se aqui a definição do AI Act da União Europeia (2024), em contraste com formulações que associam a IA à emulação de capacidades humanas, recorrentemente criticadas por seu caráter antropomorfizante, além de encobrirem os processos técnicos e interesses envolvidos.

no cálculo proposicional de Leibniz[2]. Mas foi apenas no século XX, com a invenção dos computadores, que as bases formais da IA começaram a emergir.

O britânico Alan Turing (1912-1954) foi um dos pioneiros na construção destas bases. Seu conceito de *máquina universal*[3], descrito em 1936, abriria caminho para a concepção dos primeiros sistemas que tentavam imitar a inteligência humana. Em 1943, os cientistas cognitivos McCulloch e Pitts propuseram a analogia entre neurônios biológicos e circuitos eletrônicos, introduzindo a ideia de Redes Neurais Artificiais (RNAs). Em 1946 foi criado o ENIAC, o primeiro computador eletrônico de uso geral, projetado para cálculos militares. Em 1950, nova contribuição de Turing: o cientista propôs o famoso *Teste de Turing*, experimento que avalia o sucesso da máquina ao tentar se passar por um ser humano.

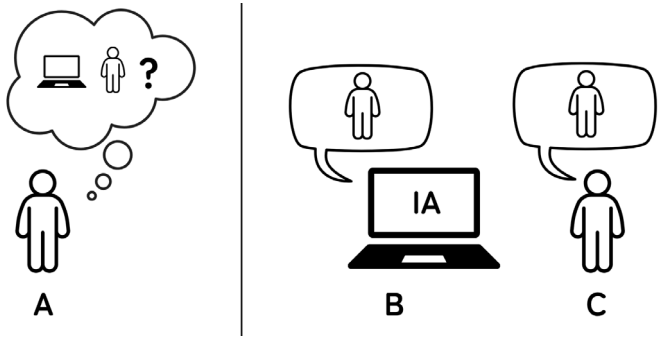

Figura 2: Representação do Teste de Turing. Um avaliador humano (A) interage via texto com uma máquina (B) e com outro humano (C), tentando identificá-los com base em suas respostas. Caso o avaliador não consiga distinguir qual é a máquina, considera-se que esta, ao demonstrar comportamento equivalente ao humano, passou no teste. Elaboração própria.

2. O cálculo proposicional é um sistema lógico que estuda proposições e suas combinações por conectivos lógicos, como "e", "ou" e "não".

3. Alan Turing propôs o modelo teórico da *máquina universal*, que seria capaz de simular qualquer outro processo computacional.

O Teste de Turing levantou questões que ainda hoje são debatidas: *Como diferenciar uma máquina consciente de uma que apenas imita comportamentos humanos? As máquinas de fato pensam ou apenas simulam o pensamento? Uma máquina pode ter consciência, livre-arbítrio ou sentimentos? E, se uma máquina se declarar consciente, ela deveria ter (quais) direitos?*

Todos esses avanços entrelaçam campos como matemática, biologia, neurociência e computação, preparando terreno para o surgimento da IA, mas sem ainda caracterizá-la como um campo autônomo.

O início da história da IA propriamente dita, enquanto campo de pesquisa autônomo, começa na década de 1950, a partir da Conferência de Dartmouth. A partir daí, sua evolução pode ser explicada em quatro fases: (I) a era dos *métodos simbólicos* (entre as décadas de 1950 e 1970); (II) o advento dos *sistemas especialistas e a evolução das RNAs* (entre os anos 1970-2000); (III) a *popularização do machine learning* a partir dos anos 2000 e, finalmente; (IV) o avanço do *deep learning* a partir dos anos 2010. Existem outros marcos e possibilidades de divisão dessa trajetória, mas o esquema aqui apresentado será suficiente para oferecer uma abordagem evolutiva clara e didática da IA.

Antes de mais nada, um aviso aos navegantes: seria um erro *delimitar rigidamente* tais fases, como se uma findasse de imediato, sendo logo substituída pela outra. Os avanços em IA processam-se de forma cumulativa e interdependente, com as fases por vezes se sobrepondo e evoluindo simultaneamente. Como já prelecionava Gilbert Simondon (1924-1989), o processo inventivo e a construção de novas abordagens não são lineares, mas refletem o contínuo refinamento e adaptação das técnicas anteriores.

Os períodos chamados "invernos da IA", são aqueles nos quais houve um menor interesse na área. Isso não quer dizer que nestes momentos as pesquisas cessaram ou, tampouco, que a área sofreu retrocessos. O que ocorreu nos invernos da IA foi, antes, uma *quebra das expectativas*: promessas não concretizadas e a ausência de descobertas impactantes induziram a um certo ceticismo, à redução do interesse do público em geral e a cortes nos financiamentos. No entanto, ainda que de forma mais modesta, as evoluções continuaram acontecendo, e foi também graças a elas que a IA pôde se reinventar e progredir.

Historicamente, a primeira fase da IA, conhecida como *simbólica*, estendeu-se entre as décadas de 1950 a 1970, sendo marcada por sistemas que utilizavam regras explícitas e símbolos para representar o raciocínio humano em contextos específicos e bem definidos. A título de ilustração: um sistema simbólico pode usar a regra "Se A, *então* B", onde A representa "está chovendo" e B significa "leve um guarda-chuva". A inteligência artificial simbólica, portanto, conecta condições (A) a ações (B) de forma lógica e estruturada.

Um conceito importante já aparece aqui de forma primitiva: o de Redes Neurais Artificiais (RNAs). Concebidas teoricamente em 1940, as RNAs são sistemas inspirados no cérebro humano, compostos por camadas de "neurônios" que processam dados e aprendem padrões. Por exemplo, o sistema Perceptron usava RNAs para aprender a separar *coisas* em dois grupos (como gatos e cachorros) a partir de amostras. As RNAs fazem isso ajustando números chamados "pesos", que ajudam a decidir de qual lado cada item analisado deve ficar, como se desenhasse uma linha imaginária entre os grupos. Cada vez que erra, o algoritmo ajusta esses pesos para melhorar na próxima tentativa.

Foi mais ou menos nessa época que surgiram os primeiros *sistemas especialistas*. Enquanto as RNAs "aprendiam por conta própria", ainda que de maneira rudimentar, os *sistemas especialistas* agiam a partir de um *repositório de conhecimento* criado por especialistas humanos, chamado "base de regras" ou "base de conhecimento". Esses sistemas eram usados para resolver problemas específicos em áreas como medicina e engenharia, funcionando como "consultores digitais" que seguiam regras já definidas para chegar a conclusões.

Mas a IA simbólica enfrentava limitações consideráveis: as RNAs ainda eram muito simples e pouco eficientes. Os sistemas especialistas, por sua vez, dependiam de regras detalhadas e rígidas, que os tornavam pouco flexíveis e ineficazes em contextos mais dinâmicos. Além disso, as restrições tecnológicas da época – especialmente quanto à capacidade de processamento e armazenamento – dificultavam a aplicação prática de modelos mais ambiciosos. Esses entraves técnicos, somados à ausência de avanços disruptivos, marcaram o início do primeiro "inverno da IA", que expôs o descompasso entre as promessas da área e as condições reais de sua implementação.

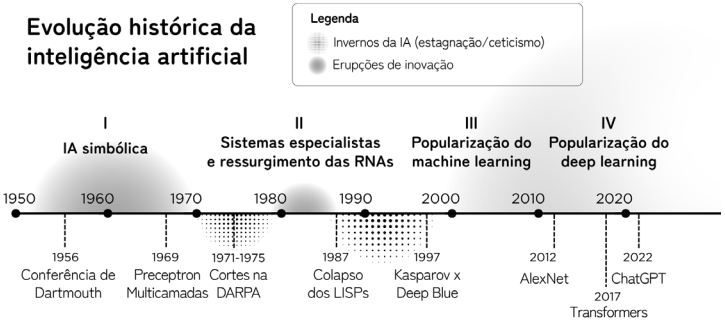

Figura 3: Fases da evolução da inteligência artificial. Elaboração própria.

O início dos anos 1970 marcou o início do primeiro "inverno" da IA. Expectativas exageradas não se concretizaram, ensejando críticas como o Relatório Lighthill (1973) no Reino Unido, que classificou muitos projetos de IA como "incapazes de gerar resultados práticos". Nos Estados Unidos, entre 1971 e 1975, a DARPA[4], principal financiadora da pesquisa em IA na década anterior, sofreu um corte orçamentário drástico, passando a exigir que os projetos demonstrassem resultados imediatos ou aplicações militares claras – o que a maioria dos sistemas ainda não conseguia fazer.

Na década de 1980, a IA ensaiaria uma retomada, graças ao advento de uma nova técnica de RNAs: a *backpropagation*[5] (do inglês: retropropagação do erro). A retropropagação refinou o ajuste de pesos das conexões neurais, reduzindo o erro e tornando os sistemas mais precisos. Essa técnica resgatou o potencial das RNAs, ampliando suas aplicações em áreas como reconhecimento de padrões e classificação de dados. Uma aplicação desta descoberta foi o uso pioneiro de RNAs para reconhecer caracteres em tarefas básicas, como diferenciar letras ou números em documentos.

Enquanto isso, os *sistemas especialistas* ganhavam notoriedade ao capturar o conhecimento humano em regras explícitas, sendo utilizados com relativo sucesso em

4. A DARPA (Defense Advanced Research Projects Agency) é uma agência do Departamento de Defesa dos Estados Unidos, fundada em 1958, que desenvolve tecnologias para a segurança nacional.

5. A técnica de *retropropagação do erro*, apresentada por Rumelhart, Hinton e Williams em 1986, usa os pesos das conexões para treinar as redes neurais. Funciona calculando o erro na saída e "voltando" por todas as camadas da rede, corrigindo os pesos para melhorar os resultados na próxima tentativa.

campos como diagnóstico médico e engenharia. Esses avanços renovaram o otimismo com a IA, que parecia caminhar para uma fase promissora. Um caso célebre foi o sistema especialista MYCIN, que ajudava médicos a identificar infecções e recomendar antibióticos, funcionando como um "doutor digital". O MYCIN funcionava fazendo perguntas aos médicos sobre os sintomas do paciente, como febre, exames ou tipo de infecção. Ele então usava essas informações para consultar sua base de regras, como "Se o paciente tiver X e Y, então é provável que seja Z", e sugeria o melhor antibiótico e a dosagem adequada. Tudo baseado no conhecimento que especialistas humanos haviam introduzido previamente no sistema.

No entanto, a euforia com a IA nos anos 80 durou pouco. O alto custo das tecnologias disponíveis, como as máquinas LISP[6], e a limitação do poder computacional da época impediram uma adoção mais ampla. Em 1987, o mercado de computadores especializados em IA (que incluía os LISPs), acabaria por colapsar ao ser superado pelas *workstations* convencionais, que eram mais baratas e eficientes. Essas dificuldades, somadas ao desânimo com a incapacidade dos sistemas de resolverem tarefas mais complexas, acabaram por conduzir a um segundo "inverno da IA". Esse período durou até meados da década de 1990, interrompendo temporariamente o entusiasmo em torno do campo.

6. As máquinas LISP eram computadores especializados nos anos 1980, usados em laboratórios e empresas para criar sistemas especialistas com a linguagem LISP. Elas resolviam problemas complexos, como diagnósticos médicos e engenharia, mas eram caras e inacessíveis para o uso pessoal. Cada computador LISP chegava a custar mais de 100 mil dólares. Com o tempo, foram substituídas por computadores mais baratos e com nível similar de desempenho, como aqueles fabricados pela Sun Microsystems.

É apenas no final da década de 1990 que a IA retoma sua força de vez. A conquista mais lembrada é o triunfo, em 1997, do supercomputador Deep Blue (IBM) sobre o maior enxadrista do mundo, Garry Kasparov. Utilizando *força bruta* de processamento e algoritmos de busca otimizados, o Deep Blue analisava milhões de posições por segundo para calcular o melhor movimento em cada jogada de xadrez.

Tecnicamente, o fato mais marcante no final do século XX foi a popularização do aprendizado de máquina (do inglês: *machine learning* – ML), subárea da IA que permite que os sistemas aprendam a partir de dados, sem necessidade de programação humana. O conceito remonta à década de 1970, mas sua popularização veio apenas no final do século, quando a evolução computacional fez com que esses sistemas se tornassem incrivelmente precisos. O ML, evolução das RNAs primitivas, começou a libertar o algoritmo da necessidade de regras predefinidas, praticamente indispensáveis para os antigos sistemas especialistas.

MACHINE LEARNING (ML)
Ramo da IA que permite que sistemas **aprendam padrões** e façam previsões com base em dados, **sem programação explícita** para cada tarefa.

Esse avanço continuou no início dos anos 2000, resultando no desenvolvimento do *deep learning* (DL). O DL emprega Redes Neurais Profundas *(Deep Neural Networks* – DNNs*),* das quais Redes Neurais Convolucionais *(Convolutional Neural Networks* – CNNs*)* e Recorrentes *(Recurrent Neural Networks* – RNNs*)* são variações especializadas para diferentes tipos de dados.

Dentre outras possibilidades, o DL favorece o processamento de dados visuais de forma eficiente, analisando imagens em partes para identificar padrões como bordas, texturas e formas. Um marco importante foi o advento da LeNet-5, que utilizava CNNs para o reconhecimento óptico de caracteres (OCR) manuscritos. Essa tecnologia foi aplicada a tarefas práticas, como a leitura de códigos postais e talões de cheques, demonstrando o potencial do DL no reconhecimento visual e abrindo caminho para inovações futuras.

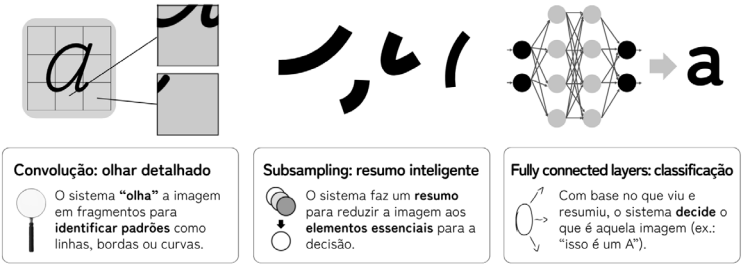

Figura 4: *deep learning* em ação para o reconhecimento de caracteres: camadas de convolução extraem padrões, camadas de subamostragem (do inglês: *subsampling*) reduzem a imagem aos elementos mais relevantes e, por fim, camadas totalmente conectadas (do inglês: *fully connected layers*) fazem a classificação. Elaboração própria.

O aprendizado profundo, ou *deep learning*, portanto, é um progresso decorrente do aprendizado de máquina, caracterizando-se pelo uso de redes neurais artificiais compostas por múltiplas camadas, projetadas para extrair e processar características complexas dos dados.

DEEP LEARNING (DL)

Subcampo do ML que usa **redes neurais profundas** para processar **dados complexos**, sendo eficaz em reconhecimento de imagens, fala e linguagem natural.

Essa abordagem permite que os sistemas executem tarefas muito avançadas, que se equiparam às habilidades de um ser humano médio, tais como a direção autônoma, o reconhecimento facial e a tradução. Tudo isso graças à capacidade algorítmica para correlacionar grandes volumes de dados não estruturados, como imagens e áudios.

O avanço do DL foi impulsionado a partir de 2010 pelo aumento do desempenho computacional, especialmente com o uso das GPUs *(Graphic Processing Units)*, que possibilitaram o treinamento de modelos complexos num período de tempo aceitável. Diferentemente dos modelos tradicionais, que analisam dados de forma menos profunda, o *deep learning* utiliza essas camadas para generalizar e tomar decisões autônomas. Isso transformou significativamente o campo da IA, dando origem, por exemplo, aos *Large Language Models* (LLMs), ou modelos de linguagem grande[7].

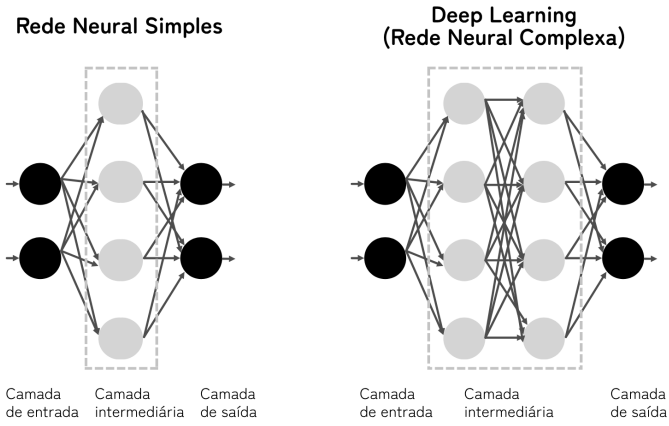

Figura 5: Diferença entre a rede neural simples e a rede neural no *deep learning*. Elaboração própria, baseada em Ryu, Chung & Choi (2021).

7. Os LLMs, são sistemas treinados para compreender e gerar texto de forma natural, a partir de grandes quantidades de dados, como o ChatGPT, o Gemini

A partir de 2017, os modelos Transformers proporcionaram um novo salto evolutivo no campo. Antes disso, os algoritmos liam palavra por palavra, como quem segue uma fila – processo lento e que limitava a compreensão de frases longas. Com os Transformers, o modelo passou a analisar todas as palavras ao mesmo tempo, mesmo quando distantes entre si. Para isso, utiliza um *mecanismo de atenção*, que se expressa por meio de uma *matriz*. Essa matriz atribui pesos às relações entre os termos, destacando os mais relevantes no contexto. Graças a essa técnica, o processamento e a compreensão dos modelos aumentaram substancialmente, possibilitando o aperfeiçoamento de tecnologias como o Google Tradutor e impulsionando o desenvolvimento dos modelos de IA generativa, capazes de gerar, resumir e interpretar textos.

IA GENERATIVA

Redes neurais treinadas com grandes volumes de dados e capazes de produzir conteúdo em linguagem natural, permitindo a realização de tarefas como escrita, tradução e geração de imagens

A maturação do *deep learning*, consolidada ao longo da década de 2010 e aprofundada nos anos seguintes, conferiu à IA robustez técnica e expressividade funcional. A combinação entre redes neurais profundas, capacidade computacional ampliada e oceanos de dados digitais permitiu um avanço significativo da inteligência artificial.

e o DeepSeek. Na atualidade, estes sistemas funcionam como um super dicionário interativo, que, em vez de apenas armazenar palavras, entende e gera frases complexas, adaptando-se ao contexto da conversa. Isso permite sua utilização como assistentes nos mais variados domínios.

Aplicações que, até pouco tempo, eram vistas como ficção científica tornaram-se realidade: reconhecimento facial em tempo real, tradução multilíngue instantânea, transcrição automatizada de fala e algoritmos capazes de diagnosticar doenças com precisão equiparável à de especialistas humanos. Tecnologias como o AlphaGo, que superou campeões mundiais no complexo jogo de tabuleiro Go, e redes convolucionais treinadas para identificar câncer de pele a partir de imagens dermatológicas, evidenciam o salto técnico ocorrido no período.

A nova geração de sistemas não depende mais de instruções explícitas programadas, mas da detecção autônoma de padrões em grandes volumes de dados, inaugurando um paradigma de *autonomia algorítmica*. Trata-se de uma transição silenciosa, porém decisiva, em que a IA deixa de ser apenas ferramenta para se tornar colaboradora – abrindo caminho para a popularização da inteligência artificial generativa.

⊡ IA generativa e a era dos agentes

A partir de novembro de 2022, com a liberação pública do ChatGPT, inicia-se uma nova etapa na história da inteligência artificial: a era da IA generativa. Com interfaces acessíveis e comandos simples *(prompts)*, modelos como Claude, Gemini, Perplexity AI e Grok permitiram aos usuários interações personalizadas – como assumir papéis específicos, analisar documentos ou estruturar argumentos complexos, inclusive com base em modelos como FIRAC[8]

8. Acrônimo do método FIRAC que oferta uma estrutura analítica para resolver problemas jurídicos, dividida em cinco etapas: Fatos (*Facts*), Questão a ser resolvida (*Issue*), Regra (*Rule*), Análise (*Analysis*), Conclusão (*Conclusion*).

ou Toulmin[9]. Essa evolução deu origem aos assistentes generativos, capazes de desempenhar tarefas especializadas com agilidade e adaptabilidade, como resumir reuniões, redigir pareceres ou organizar pesquisas jurídicas.

Mais recentemente, o desenvolvimento de agentes autônomos – como Tasks e Operator, da OpenAI – marcou uma nova virada. Diferentemente dos modelos geradores de conteúdo sob demanda, os agentes são capazes de atuar de forma proativa e autônoma, coordenando ações em ecossistemas digitais complexos. Conforme Noam Kolt, enquanto os LLMs operam como "copilotos", os agentes se comportam como "pilotos automáticos", decidindo e agindo para atingir objetivos definidos pelo usuário. Embora ainda incipiente no campo jurídico, esse paradigma técnico já demonstra potencial em setores como logística, consumo, educação e saúde.

Em termos conceituais, *agentes* são sistemas que recebem objetivos gerais e decidem como cumpri-los com base em informações ambientais e memória operacional. Segundo Bornet *et al.* (2025), sua arquitetura se apoia em três capacidades fundamentais: ação, raciocínio e memória. A ação refere-se à execução de comandos em ambientes digitais; o raciocínio, à análise de cenários e tomada de decisões; e a memória, à aprendizagem com interações anteriores, permitindo ajustes estratégicos e consistência comportamental.

Os agentes inauguram uma nova fase da IA, marcada por maior autonomia e menor supervisão humana. Já são

9. O filósofo britânico Stephen Toulmin (1922-2009) propôs um modelo que estrutura a argumentação em seis elementos: alegação, fundamento, justificativa, qualificador, evidência de apoio e refutação.

utilizados para apoiar pessoas com deficiência, coordenar emergências e realizar tarefas com impacto jurídico direto, como contratações e administração de recursos. No entanto, sua atuação levanta sérios riscos de segurança, responsabilidade e alinhamento ético. Apesar do alto impacto potencial, menos de 20% dos agentes possuem políticas formais de segurança, e apenas 10% passaram por auditorias externas, evidenciando um déficit preocupante de governança. A situação se torna ainda mais complexa com a interação entre múltiplos agentes, que pode desencadear ações adversas automatizadas.

Diante disso, torna-se urgente repensar os marcos regulatórios e teóricos aplicáveis. Os agentes desafiam categorias jurídicas tradicionais como vontade, culpa e intenção, exigindo uma reconstrução conceitual da responsabilidade. Para enfrentar esse cenário, Bornet *et al.* (2025) propõem uma governança híbrida, baseada na colaboração entre humanos e agentes e estruturada em cinco pilares: rastreabilidade, explicabilidade, controle contínuo, sistemas de feedback e documentação técnica. Esse modelo distribui deveres entre desenvolvedores, operadores, reguladores e instituições, inspirando-se na lógica da responsabilidade por produtos e das redes contratuais. O desafio não é apenas normativo, mas conceitual: como atribuir responsabilidade a entidades que agem, mas não possuem vontade própria? Ignorar essa questão poderá gerar consequências jurídicas e sociais significativas.

À medida que avançamos para o futuro, a IA promete transformar ainda mais profundamente a sociedade. Neste ponto, faz-se necessário demarcar alguns conceitos. Conforme as normas internacionais estabelecidas pela ISO/IEC (2022), a IA fraca *(weak AI)* é aquela que processa