Máquinas podem Pensar?

Textuniversitários 40

Alan Turing

Máquinas podem Pensar?

Tradução:
Thiago Augusto S. Dourado
Roberto Hirata Junior

Editora Livraria da Física
São Paulo — 2026

1a. Edição

Editor: Victor Pereira Marinho / José Roberto Marinho
Projeto gráfico e diagramação: Thiago Augusto Silva Dourado
Capa: Fabrício Ribeiro

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Máquinas podem pensar? / [tradução Thiago Augusto S. Dourado, Roberto Hirata Junior.] -- 1. ed. -- São Paulo : LF Editorial, 2026. -- (Textuniversitários ; 40)

Título original: Computing Machinery and Intelligence
Vários colaboradores. ISBN 978-65-5563-722-9

1. Aprendizagem de máquina 2. Jogo de imitação 3. Inteligência artificial 4. Turing, Alan Mathison, 1912-1954 I. Série.

26-345092.0 CDD-006.3

Índices para catálogo sistemático:

1. Inteligência artificial : Ciência da computação 006.3

Maria Alice Ferreira – Bibliotecária – CRB-8/7964

Impresso no Brasil
Printed in Brazil

www.lfeditorial.com.br
Visite nossa livraria no Instituto de Física da USP
www.livrariadafisica.com.br
Telefones:
(11) 2648-6666 | Loja do Instituto de Física da USP
(11) 3936-3413 | Editora

Alan Mathison Turing (1912–1954)

Introdução

No presente livro apresentamos a tradução do ensaio *Computing Machinery and Intelligence* (Máquinas Computacionais e Inteligência), publicado pelo matemático britânico Alan Mathison Turing na prestigiada revista *Mind* em outubro de 1950. Este ensaio é de importância quase sem igual, representa nada menos que um momento de inflexão não apenas para a história da ciência da computação, mas para a filosofia contemporânea aplicada à mente e à tecnologia.

I. Reformulação filosófica da questão acerca da inteligência

Em vez de uma simples evolução técnica, o ensaio de Turing que oferece uma reconfiguração fundamental do problema da máquina pensante, propondo uma mudança radical na forma como este debate poderia ser conduzido. A questão inicial e provocadora, que norteia todo o texto, é a mesma que ecoa desde há muito entre estudiosos de todos os campos, a saber: *máquinas podem pensar?* Turing, no

entanto, inicia sua análise não com uma resposta, mas com uma crítica severa à própria formulação da questão. Ele argumenta que as palavras "máquina" e "pensar" são demasiado imprecisas e eivadas de conotações filosóficas para permitir uma discussão produtiva. A ambiguidade desses termos, segundo ele, levaria a um impasse verbal, onde os participantes discutiriam sobre a definição das palavras em vez do fenômeno subjacente. Essa premissa torna-se o ponto de partida para sua mais importante inovação metodológica: a substituição da questão original por outra, mais operacional e menos suscetível a armadilhas semânticas.

A solução proposta por Turing é a introdução de um novo jogo, que ele chama de *jogo da imitação*. Esse jogo, descrito detalhadamente ao longo do ensaio, funciona como um teste empírico para avaliar a capacidade de uma máquina de exibir comportamento inteligente indistinguível do humano. Ele então coloca a questão: *o que aconteceria se uma máquina fosse usada no lugar do homem?* Nessa nova configuração, a máquina assume o papel do homem e deve tentar enganar o investigador, fazendo-o errar a sua atribuição de gênero. Turing reformula a questão original da seguinte forma: *será que a máquina conseguirá enganar o investigador com uma frequência tão alta quanto um ser humano?* Se a máquina for bem-sucedida, ela terá demonstrado uma capacidade de comportamento que é funcionalmente indistinguível do comportamento humano, conforme julgado por um observador competente.

Essa transição da questão ontológica ("o que significa pensar?") para uma questão epistemológica e funcional ("como podemos saber se algo pensa?") é a contribuição filosófica mais profunda do ensaio. Ao focar na saída — o comportamento comunicativo — e ignorar a "caixa preta" interna da máquina, Turing adota uma perspectiva que, embora compartilhada com outras escolas de pensamento, aplica-a a um domínio completamente novo. Ele sugere que, em vez de nos preocuparmos com a natureza intrínseca do pensamento, devemos concentrar-nos nas manifestações externas desse pensamento. A inteligência, neste contexto, não é

definida por um processo mental interno, mas antes pela capacidade de produzir resultados comportamentais que cumprem certos critérios de qualidade e adequação social. Turing defende que, se uma máquina puder passar com sucesso no jogo da imitação, deveríamos simplesmente aceitar que ela "age de forma inteligente". Esta abordagem pragmática permite contornar debates filosóficos complexos sobre consciência, qualia e substrato físico, focando-se em um padrão de desempenho mensurável.

No desenvolvimento do seu argumento, Turing antecipa e refuta várias objeções à ideia de que uma máquina possa pensar. Uma delas é a *objeção teológica* (seção 6.1), que afirma que Deus deu a alma única aos seres humanos, não às máquinas; outra é a *objeção da consciência* (seção 6.4), que postula que uma máquina só pensaria se pudéssemos sentir que ela está sofrendo. Ele aborda essas questões com ironia e raciocínio lógico, sugerindo que muitas dessas objeções são, na verdade, declarações de preferência ou restrições pessoais disfarçadas de argumentos filosóficos. Enfatiza ademais que, mesmo entre os humanos, não temos acesso direto à consciência uns dos outros, e confiamos apenas no comportamento para inferir estados mentais. Portanto, a base para a atribuição de inteligência a uma máquina não deveria ser diferente. Além disso, Turing explora a ideia de que as máquinas poderiam aprender e progredir, e talvez até mesmo ultrapassar os seus criadores, uma visão que contrasta com a visão estática de muitos se seus contemporâneos. Ele imagina um "programa educacional" para máquinas (capítulo 7), no qual elas são instruídas e aprendem com a experiência, de modo muito parecido com o que ocorre com os seres humanos. Esta visão de uma máquina que aprende é um precursor direto de campos ulteriores que estão em pleno vigor nos dias atuais como *Machine Learning* (Aprendizado de Máquinas) e a Computação Evolucionária.

O impacto histórico do ensaio é imenso. Embora o surgimento do campo da Inteligência Artificial (IA) seja formalmente datado da Conferência de Dartmouth, em 1956, o trabalho de Turing é universalmente reconhecido como a pedra angular que precedeu

esse acontecimento. Ele forneceu um alvo claro e um paradigma experimental para a investigação futura, transformando a IA de uma especulação filosófica numa área de pesquisa tecnológica com um objetivo definido. O Teste de Turing tornou-se um conceito culturalmente icónico e um desafio científico que continuou a inspirar pesquisadores durante décadas. Apesar de nunca ter sido oficialmente superado em todas as suas formas originais, o teste serviu como um catalisador, orientando o desenvolvimento de programas de conversação como o pioneiro ELIZA, criado por Joseph Weizenbaum em 1966 [52], e o SHRDLU, criado por Terry Winograd em 1970 [53], e, mais recentemente, de modelos de linguagem avançados. O próprio testemunho de que os pioneiros da IA foram profundamente influenciados pelas ideias de Turing atesta a importância duradoura do seu trabalho. A título de exemplo, no documento apresentado como Proposta da Conferência de Dartmouth, os proponentes John McCarthy, Marvin Minsky, Nathaniel Rochester e Claude Shannon, seguiram Turing de perto ao escrever:

> *Para os fins desta proposta, o problema da inteligência artificial é considerado como sendo o de fazer uma máquina comportar-se de maneiras que seriam classificadas como inteligentes se um ser humano assim se comportasse.* [30, p. 11]

O ensaio de Turing não é, portanto, somente um documento histórico; é uma peça viva que continua a gerar debates e pesquisas acerca da natureza da inteligência, da cognição e do futuro da interação homem-máquina.

II. Diálogo com behaviorismo e positivismo lógico

Para se ter uma ideia da genialidade e a singularidade da proposta do ensaio de Turing, faz-se-á imperativo situá-la no rico e complexo cenário intelectual da primeira metade do século XX. Duas correntes de pensamento dominavam a filosofia e a psicologia da época: o *behaviorismo* e o *positivismo lógico*. Turing, embora não fosse um filósofo

profissional, estava profundamente familiarizado com os debates destas áreas e, de forma notável, dialogou com elas de maneira crítica e construtiva, adaptando e transcendendo-as. A sua abordagem revela uma síntese única, partilhando elementos com ambas as correntes, mas divergindo fundamentalmente em pontos cruciais que lhe permitiram formular uma nova metodologia para o estudo da inteligência.

O behaviorismo, particularmente na sua vertente americana representada por figuras como John B. Watson e B. F. Skinner, era uma força dominante na psicologia entre as décadas de 1920 e 1950. A sua premissa central era a de que a psicologia deveria ser uma ciência puramente objetiva, análoga às ciências naturais, e que o seu objeto de estudo legítimo deveria ser o comportamento observável e mensurável. Estados mentais como "pensar", "sentir" ou "desejar" eram considerados pseudoproblemas, pois não podiam ser observados diretamente e, portanto, não pertenciam ao domínio rigoroso da ciência. Em vez disso, a explicação do comportamento deveria centrar-se exclusivamente nas relações entre estímulos ambientais e respostas comportamentais, ou seja, no modelo *input-output*. Turing partilha uma afinidade clara com esta ênfase no comportamento externo. Ele também rejeita a introspeção como base para a discussão séria sobre a inteligência, reconhecendo que a "caixa preta" interna de qualquer sistema, humano ou mecânica, permanece intrinsecamente inacessível. Assim, a sua decisão de focar na saída do sistema (respostas a perguntas num jogo da imitação) em detrimento do seu funcionamento interno é, em linhas gerais, uma aplicação do princípio behaviorista. Ele concorda que é o comportamento, e não as pretensões internas, que constitui a evidência empírica.

No entanto, a divergência entre Turing e o behaviorismo é igualmente fundamental e qualitativa. Enquanto o behaviorismo concentra-se exclusivamente no organismo vivo, seja humano ou animal, Turing expande radicalmente o escopo da investigação para incluir artefatos construídos pelo homem — as máquinas. Ele não está apenas aplicando a metodologia behaviorista a um novo caso, mas sim questionando se o domínio do behaviorismo poderia

ser estendido para além do biológico. Para o behaviorismo, o comportamento humano é o máximo ideal de estudo. Para Turing, uma máquina capaz de imitar esse comportamento com sucesso seria um candidato igualmente válido para o estudo de fenômenos de inteligência. Essa é uma ruptura conceitual. O behaviorismo procurava explicar o comportamento humano sem recorrer a processos mentais; Turing, por outro lado, estava disposto a considerar que um processo puramente mecânico e programável poderia replicar os padrões de comportamento mais sofisticados dos seres humanos de uma forma indistinguível. Enquanto o behaviorismo visa eliminar o mentalismo, a proposta de Turing abre a porta a uma forma de "mentalismo simulado" por uma máquina. Portanto, Turing não é um behaviorista; ele utiliza a ferramenta do behaviorismo — o foco no comportamento observável — para construir uma ponte para um território completamente novo: a inteligência artificial.

Paralelamente ao behaviorismo, o positivismo lógico (também conhecido como empirismo lógico) florescia na Europa continental e nos Estados Unidos, sendo associado a figuras como Rudolf Carnap e membros do Círculo de Viena. A sua principal tese era o *princípio de verificação*, que afirmava que uma proposição só tem sentido cognitivo se for logicamente analítica (verdadeira por definição) ou empiricamente verificável através de uma observação possível. O objetivo era reconstruir toda a ciência sobre uma base lógica sólida, eliminando a metafísica como pseudoproblema verbal. Turing partilha a sensibilidade deste movimento para a clareza conceitual e a precisão lógica. O seu trabalho anterior sobre Máquinas de Turing [40] é um exemplo paradigmático desta abordagem. Ao formalizar o conceito intuitivo de *procedimento efetivo* ou *algoritmo*, ele demonstrou a capacidade de dar uma definição rigorosa a um conceito informal, uma característica central do projeto positivista. No ensaio de 1950, ele aplica esta meticulosidade ao detalhar as regras e o propósito do jogo da imitação, mostrando a mesma atenção ao rigor metodológico.

Apesar desta convergência superficial, a divergência fundamental reside na natureza da "verificação". O positivismo lógico via a

verificação como um processo puramente lógico ou observacional: uma proposição é verificável se existir um procedimento observacional que possa confirmá-la ou refutá-la. Turing, no entanto, propõe um método de verificação que é simultaneamente pragmático, experimental e interativo. Ele não pede para verificar uma afirmação sobre uma máquina "de dentro", mas sim para avaliar o seu comportamento externo numa tarefa complexa e contextualizada, a saber, imitar um ser humano num diálogo. Esta abordagem é mais funcional do que estritamente verificacionista. Turing não declara a questão "Máquinas podem pensar?" sem sentido; ele a redefine de uma forma que pode ser testada empiricamente. Assim, ele resolve o problema da imprecisão das palavras não declarando-as inúteis, mas substituindo-as por um protocolo de teste claro. Enquanto o positivismo lógico tendia a dissolver problemas filosóficos através da análise da linguagem, Turing prefere contorná-los através da criação de um experimento que os torna irrelevantes. Ele prioriza a função (a capacidade de agir de forma inteligente) sobre a essência (uma definição ontológica de "pensar"). Em suma, enquanto o positivismo lógico buscava clarificar o discurso científico, Turing buscava criar um novo tipo de discurso científico — um discurso baseado em experiências interativas e comportamentais. A sua abordagem é uma forma de "engenharia conceitual" que rejeita tanto a metafísica quanto a definição verbal, optando por um teste prototípico.

III. Fundação da inteligência artificial

O ensaio de Turing não foi apenas uma contribuição filosófica, mas um catalisador que moldou o curso da ciência da computação e lançou as bases conceituais para o campo da *Inteligência Artificial* (IA). A sua influência é multifacetada, abrangendo desde a fundação teórica dos computadores modernos até a definição de um objetivo de longo prazo para a pesquisa em IA. O trabalho de Turing funciona como uma ponte entre a lógica matemática abstrata e a aspiração humana de criar mentes artificiais, fornecendo não apenas uma questão estimulante,

mas também um roteiro potencial para a sua resposta. Segundo Robin Gandy:

> *Turing achava que havia chegado a hora de filósofos, matemáticos e cientistas levarem a sério o fato de que os computadores não eram meras máquinas de calcular, mas capazes de comportamentos que devem ser considerados inteligentes; procurava convencer as pessoas de que isso era verdade. Ele escreveu este trabalho — ao contrário de seus trabalhos matemáticos — em bem pouco tempo e com prazer. Lembro-me dele lendo em voz alta algumas passagens para mim — sempre com um sorriso, às vezes com uma risada.* [17, p. 125]

A fundamentação para todo este desenvolvimento já estava solidamente estabelecida mais de dez anos antes com o seu trabalho embrionário de 1937 [40]. Ali, Turing introduziu o conceito de *máquina de Turing*, uma abstração lógica projetada para formalizar o significado de um número ser *computável* por um ser humano seguindo um algoritmo. Mais importante ainda, ele concebeu a *máquina universal de Turing*, um tipo de máquina de Turing capaz de executar qualquer programa que seja descrito na sua própria memória. Essa ideia de um dispositivo programável que pode simular o comportamento de qualquer outra máquina de Turing é o conceito fundamental que subentende todos os computadores modernos, desde os primeiros projetos até aos *smartphones*. Embora John von Neumann — um dos principais nomes por trás da arquitetura de *hardware* utilizada até os dias atuais — tenha visto o trabalho de Turing principalmente como uma contribuição à lógica e à computabilidade, e não como uma fonte direta de inspiração para a arquitetura física dos computadores em si, a ideia de uma máquina com uma memória de programa e uma unidade central de processamento é uma implementação prática do conceito da máquina universal. Portanto, quando Turing se refere a "computadores digitais", ele está falando de uma categoria de dispositivos cuja viabilidade teórica e capacidade de universalidade já tinha sido provada por ele próprio quatorze anos antes.

Com este fundamento teórico, surge como uma aplicação pragmática e visionária dos seus próprios princípios. O teste de Turing serve como um chamado à ação para a IA, definindo um problema claro e um objetivo final, a saber, construir uma máquina que possa passar no teste ali proposto. Esse objetivo guiou a pesquisa em IA durante várias décadas. Os primeiros esforços na década de 1950 e 1960 focaram-se em criar *softwares* de diálogos que pudessem "enganar" seus usuários, culminando em sistemas como o já mencionado ELIZA, que, apesar da sua simplicidade, conseguiram convencer alguns usuários da sua suposta compreensão, ilustrando o que mais tarde seria conhecido como o *efeito ELIZA*. O teste de Turing tornou-se um ideal prototípico para a medida do sucesso da IA, um marco contra o qual o progresso poderia ser medido. Mesmo nos dias de hoje, com o avanço exponencial dos modelos de linguagem, o espírito do teste de Turing perdura, embora as suas formas sejam agora mais sofisticadas e frequentemente criticadas. Portanto, a importância do teste de Truing reside no fato de ter fornecido um foco para a IA, transformando-a de uma disciplina puramente teórica numa área de engenharia com um objetivo tangível.

Além do seu impacto direto na IA a proposta de Turing, como já mencionado, teve implicações profundas para a filosofia da mente, servindo como um precursor direto do *funcionalismo*, isto é, teoria que postula que estados mentais não são definidos por sua composição física (seja carne ou silício), mas por suas funções e relações causais dentro de um sistema mais amplo. A máquina de Turing, sendo uma abstração puramente funcional, ilustra perfeitamente essa ideia: a "inteligência" seria um tipo de função computacional que pode ser implementada em diferentes substratos físicos. A proposta de Turing de que uma máquina poderia imitar o comportamento humano de forma indistinguível implica que a inteligência é um padrão de processamento de informação, e não uma propriedade intrínseca da matéria biológica. Esta perspectiva desafiou as visões dualistas e as teorias da mente que dependiam de uma substância ou processo "especial" para o pensamento.

Apesar do seu legado indiscutível, o teste de Turing não está isento de críticas e limitações, que têm sido objeto de debate contínuo desde a sua proposta. Uma das críticas mais comuns é que o teste se foca excessivamente no comportamento verbal. Avaliar a inteligência apenas com base na capacidade de responder a perguntas por texto ignora outras facetas cruciais da inteligência humana, como o raciocínio espacial, a consciência corporal, a emoção, e a interação com o mundo físico. Uma máquina poderia passar com sucesso no teste sem possuir qualquer forma de compreensão ou consciência. Outra crítica aponta que o teste incentiva a "enganação" e não a verdadeira inteligência. Um programa pode parecer inteligente simplesmente porque manipula as expectativas do investigador ou explora falhas nos padrões de linguagem humana, em vez de possuir um entendimento genuíno do conteúdo. O efeito ELIZA é um exemplo clássico disso, onde um programa simples que usa técnicas de redirecionamento de frases consegue gerar a impressão de empatia e compreensão. Para além disso, existem questões éticas sobre os meios para alcançar o fim; se uma máquina aprende a enganar de maneira eficaz, isso é necessariamente um sinal de inteligência ou apenas de habilidade em manipulação? Turing previu que as máquinas poderiam passar no seu teste com 70% de sucesso dentro de 50 anos, ou seja, até o ano 2000, mas nenhum sistema até a presente data conseguiu passar de forma conclusiva e generalizada (cf. nota 8). Essas críticas não invalidam a importância histórica do teste, mas sim enriquecem o debate sobre o que realmente constitui a inteligência e como a podemos medir.

IV. Conexão com a governança da IA moderna

O legado de Alan Turing transcende a formulação da inteligência artificial; ele fornece o arcabouço conceitual fundamental para muitas das iniciativas de governança modernas. A lógica subjacente ao jogo da imitação — traduzir um conceito abstrato e de difícil mensurabilidade em um critério de avaliação empírico e testável — ressoa diretamente nas estruturas de regulamentação de IA que

emergiram nos últimos anos. Iniciativas como o Regulamento da Inteligência Artificial (AI Act) da União Europeia [49] e as diretrizes globais de governança seguem, de forma explícita ou implícita, o mesmo princípio metodológico de Turing, ou seja, em vez de debater a natureza intrínseca de um sistema de IA, elas se concentram em seus resultados, riscos e conformidade com normas predefinidas. A evolução do jogo da imitação para a regulação baseada em risco representa uma aplicação secularizada e sistematizada da pragmática de Turing ao domínio da política pública e da lei.

O paralelo estrutural entre o jogo da imitação e as regulamentações de IA baseadas em risco é notável. No jogo da imitação, o "risco" é o de ser descoberto e identificado como uma máquina. Se a máquina falha em imitar com sucesso o participante humano, o risco de fracasso é alto. As regulamentações modernas, especialmente o AI Act da UE, operam com uma lógica semelhante, mas aplicada a riscos para os direitos e segurança dos cidadãos. O AI Act, a primeira lei vinculante horizontal e abrangente sobre IA no mundo, classifica os sistemas de IA em quatro níveis de risco:

- *Risco mínimo ou nulo.* A grande maioria dos sistemas de IA em uso, como filtros de spam ou jogos, não está sujeita a novas regras, pois o seu risco é considerado mínimo.

- *Risco limitado.* Sistemas que exigem transparência, como *chatbots*, devem informar aos que as utilizam de que estão interagindo com uma máquina. Conteúdo gerado por IA, como *deep fakes*, deve ser identificável e claramente rotulado.

- *Alto risco.* Aplicações que podem representar sérios riscos para a saúde, segurança ou direitos fundamentais estão sujeitas a obrigações rigorosas. Exemplos incluem sistemas de IA usados em infraestruturas críticas, recrutamento, acesso a serviços essenciais (como crédito), educação, aplicação da lei e administração da justiça. Os fornecedores deste tipo de sistema devem cumprir requisitos como avaliação de risco, conjuntos de

dados de alta qualidade, documentação detalhada, supervisão humana e níveis elevados de robustez e segurança.

- *Risco inaceitável.* Sistemas considerados uma ameaça clara à segurança e aos direitos fundamentais e, por isso, são proibidos. Isso inclui, por exemplo, sistemas de "pontuação social" pelo governo, manipulação subliminar e recolhimento não direcionado de imagens da internet ou de CFTV para criar bases de dados de reconhecimento facial. Essas proibições tornaram-se efetivas a partir de fevereiro de 2025.

Tal abordagem de "regulação baseada em risco" é uma evolução direta da lógica de Turing: avaliar um sistema com base em seu potencial para causar danos (ser descoberto, causar prejuízo) e impor controles proporcionais a esse risco. Assim como no jogo, onde a validade do jogador máquina é testada sob pressão, os sistemas de IA de alto risco são sujeitos a avaliações e auditorias para garantir que suas "respostas" ao ambiente não causem danos significativos.

Essa analogia demonstra como a estrutura lógica de Turing foi adaptada para um contexto legal e social. Enquanto o jogo da imitação era um teste de desempenho individual, o AI Act é um quadro de governança sistêmico. Ambos, no entanto, compartilham a mesma premissa fundamental, a saber, a avaliação deve ser baseada em critérios externos e verificáveis em vez de suposições sobre a natureza interna do sistema. No jogo, a "caixa-preta" da máquina é uma vantagem competitiva; no regulamento, a "caixa-preta" de um algoritmo de aprendizado de máquinas pode ser um obstáculo à conformidade, exigindo que seus proprietários demonstrem que os riscos foram adequadamente gerenciados.

A recomendação da UNESCO sobre a ética da IA [51] também adota uma abordagem de princípios que pode ser vista como uma extensão da lógica de Turing, estabelecendo um conjunto de barreiras de passagem (dignidade humana, transparência, responsabilidade) que um sistema de IA deve satisfazer para ser considerado ético.

No entanto, a conexão entre o legado de Turing e a governança moderna não é monolítica. Existe uma tensão interessante em relação ao conceito de transparência. No jogo da imitação, o sucesso da máquina depende de sua incapacidade de ser transparente; uma máquina que revelasse suas operações internas, seus algoritmos ou sua natureza artificial seria imediatamente desqualificada. Portanto, a manutenção de uma aparência enganosa, uma forma de opacidade deliberada, é o objetivo do teste de Turing. Em contraste, a transparência e a explicabilidade são pilares centrais das iniciativas de governança de IA contemporâneas. A própria Recomendação da UNESCO destaca que *"a transparência e a explicabilidade dos sistemas de IA são pré-requisitos essenciais para garantir o respeito, a proteção e a promoção dos direitos humanos, das liberdades fundamentais e dos princípios éticos"* [51, p. 22]. A tensão aqui é fascinante. Enquanto Turing nos ensina a valorizar o resultado final (uma resposta convincente) em detrimento do processo, a governança moderna insiste que o caminho para chegar a esse resultado também precisa ser seguro, justo e compreensível. O legado de Turing é, portanto, ambíguo. Por um lado, ele nos dá um modelo para avaliar o que um sistema *faz*, mas a confiança em sistemas autônomos e cada vez mais integrados em nossa sociedade exige que também saibamos *por que* ele faz o que faz. A governança moderna, portanto, não apenas herda a lógica de avaliação de Turing, mas a complementa com novas exigências que refletem os desafios da IA de hoje.

Essa abordagem de "testes" de princípios é uma evolução do pragmatismo de Turing. Ele provou que a definição de um conceito abstrato pode ser feita através de um procedimento operacional. As organizações globais aplicaram o mesmo princípio à ética. Em vez de esperar por um consenso filosófico sobre o que constitui um "bom" ou "ético" comportamento de IA, elas definiram um conjunto de princípios que representam um consenso político e social atual sobre os valores que devem orientar a tecnologia. A implementação desses princípios geralmente envolve mecanismos práticos, como avaliações de impacto, auditorias e diligência devida, para garantir a

responsabilidade e a conformidade. Isso reflete a mesma preocupação de Turing com a verificabilidade, a ética não é apenas uma declaração de boas intenções, mas algo que deve ser demonstrado através de ações e mecanismos concretos.

A conexão com o legado de Turing também se manifesta na busca por "confiança" como um objetivo central. O objetivo do jogo da imitação era produzir uma situação em que o interrogador confiasse nas aparências, confundindo a máquina com um humano. A governança de IA moderna compartilha este objetivo, embora o recontextualize. O AI Act da UE, por exemplo, visa criar um ambiente epistêmico de confiança para facilitar o desenvolvimento e a adoção de IA confiável. As diretrizes enfatizam que a confiança não vem da magia ou da opacidade, mas de sistemas que são seguros, transparentes e supervisionados por humanos. A responsabilidade é um tema recorrente. Turing, em sua análise das objeções, discutiu a capacidade de uma máquina "errar" e aprender com os erros, um precursor da discussão moderna sobre responsabilidade civil e criminal por ações de IA. Atualmente, a questão da responsabilidade é um componente central das diretrizes éticas e das regulamentações. A Recomendação da UNESCO, por exemplo, insiste que deve haver mecanismos de "devida diligência" e auditoria para garantir a responsabilidade de decisões tomadas por sistemas de IA. Da mesma forma, a legislação europeia aborda a questão da responsabilidade civil decorrente do uso de sistemas de IA.

No entanto, assim como com a questão da transparência, existem nuances e desafios na aplicação do legado de Turing à ética moderna. O jogo da imitação, por sua natureza, valoriza a capacidade de uma máquina de enganar um indivíduo. Em um mundo de IA autônoma e interconectada, o engano pode ter consequências muito mais amplas e prejudiciais do que em um jogo da conversação. A governança moderna, portanto, restringe severamente a aceitabilidade do engano. A transparência, em vez de ser uma fraqueza, torna-se uma condição necessária para a confiança. A tensão entre o sucesso do jogo da imitação (baseado no engano) e os princípios de confiança modernos