A ÉTICA DA INTELIGÊNCIA ARTIFICIAL

PRINCÍPIOS, DESAFIOS E OPORTUNIDADES

LUCIANO FLORIDI

PUCPRESS

ÍNDICE

PARTE UM

ENTENDENDO A IA

A primeira parte do livro pode ser lida como uma breve introdução filosófica ao passado, presente e futuro da inteligência artificial (IA). Consiste em três capítulos. Juntos, eles fornecem a estrutura conceitual necessária para entender a segunda parte do livro, que aborda algumas questões éticas urgentes levantadas pela IA. No Capítulo 1, reconstruo a ascensão da IA no passado, não histórica ou tecnologicamente, mas conceitualmente e em termos das transformações que levaram aos sistemas de IA em uso atualmente. No Capítulo 2, articulo uma interpretação da IA contemporânea em termos de um *reservatório de agência* possibilitado por dois fatores: um divórcio entre (a) a capacidade de resolver problemas e concluir tarefas para atingir uma meta e (b) a necessidade de ser inteligente ao fazê-lo; e a transformação progressiva de nossos ambientes em uma infosfera amigável para a IA. Esse último fator torna o divórcio não apenas possível, mas bem-sucedido. No Capítulo 3, concluo a primeira parte do volume analisando os desenvolvimentos plausíveis da IA em breve, mais uma vez, não técnica ou tecnologicamente, mas conceitualmente e em termos dos tipos preferenciais de dados necessários e dos tipos de problemas mais facilmente solucionáveis pela IA.

1

PASSADO

A ascensão da IA

1.0. **Resumo**

A seção 1.1 começa oferecendo uma breve visão geral de como os desenvolvimentos digitais levaram à atual disponibilidade e ao sucesso dos sistemas de IA. A seção 1.2 interpreta o impacto disruptivo das tecnologias, ciências, práticas, produtos e serviços digitais, em resumo, *o digital*, como sendo devido à sua capacidade de copiar e colar realidades e ideias que herdamos da modernidade. Chamo isso de *cleaving power* do digital. Ilustro esse *cleaving power* por meio de alguns exemplos concretos. Em seguida, eu o utilizo para interpretar a IA como uma nova forma de *agência inteligente* provocada pela desconexão digital da agência e da inteligência, um fenômeno sem precedentes que causou algumas distrações e mal-entendidos, como "a Singularidade", por exemplo. A seção 1.3 apresenta uma breve digressão sobre a agência política, o outro tipo significativo de agência transformado pelo *cleaving power* do digital. Ela explica brevemente por que esse tópico é essencial e altamente relevante, mas também está além do escopo deste livro. A seção 1.4 retorna à questão principal de uma interpretação conceitual da IA e introduz o Capítulo 2, lembrando o leitor da dificuldade de definir e caracterizar precisamente o que é IA. Na seção final, defendo que o *design* é a contrapartida do *cleaving power* do digital e antecipo alguns dos tópicos discutidos na segunda metade do livro.

1.1. **Introdução: a revolução digital e a IA**

Em 1964, a Paramount Pictures distribuiu *Robinson Crusoé em Marte*. O filme descrevia as aventuras do Comandante Christopher 'Kit' Draper (Paul Mantee), um astronauta americano que naufragou em Marte. Passe apenas alguns minutos assistindo ao filme no YouTube e você verá como o mundo mudou radicalmente em apenas poucas décadas. Particularmente, o computador no início do filme parece um motor vitoriano com alavancas, engrenagens e mostradores, uma peça de arqueologia que o Dr. Fran-

kenstein poderia ter usado. No entanto, no final da história, Friday (Victor Lundin) é rastreado por uma espaçonave alienígena por meio pulseiras, um elemento de futurologia que parece de uma previsão angustiante.

Robinson Crusoé em Marte pertenceu a uma época diferente, tecnológica e culturalmente mais próxima do século anterior do que do nosso. Descreve uma realidade *moderna*, não contemporânea, baseada em *hardware*, não em *software*. *Laptops*, Internet, serviços da *web*, telas sensíveis ao toque, *smartphones*, relógios inteligentes, redes sociais, compras *online*, *streaming* de vídeo e música, carros sem motorista, cortadores de grama robóticos, assistentes virtuais e o Metaverso ainda estavam por vir. A IA era principalmente um projeto, não uma realidade. O filme mostra tecnologia feita de porcas e parafusos e mecanismos que seguem as desajeitadas leis da física newtoniana. Era uma realidade totalmente analógica baseada em átomos e não em *bytes*. Essa é uma realidade que a geração Y nunca experimentou, porque nasceu após o início da década de 1980. Para eles, um mundo sem tecnologias digitais é como era para mim um mundo sem carros (coincidentemente, nasci em 1964): algo de que eu tinha ouvido falar pela minha avó.

Costuma-se dizer que um *smartphone* tem em poucos centímetros muito mais capacidade de processamento do que a NASA conseguiu reunir quando Neil Armstrong aterrissou na Lua cinco anos depois de *Robinson Crusoé em Marte*, em 1969. Temos todo esse poder a um custo quase insignificante. Para o quinquagésimo aniversário do pouso na Lua em 2019, muitos artigos fizeram comparações e aqui estão alguns fatos surpreendentes. O computador de orientação da Apollo (*Apollo Guidance Computer* ou AGC) a bordo da Apollo 11 tinha 32.768 *bits* de memória RAM e 589.824 *bits* (72 KB) de memória ROM. Você não poderia guardar este livro nele. Cinquenta anos depois, um telefone comum vinha com 4 GB de RAM e 512 GB de ROM. Isso representa cerca de 1 milhão de vezes mais RAM e 7 milhões de vezes mais ROM. Quanto ao processador, o AGC funcionava a 0,043 MHz. Diz-se que o processador médio de um iPhone funciona a cerca de 2490 MHz, o que é cerca de 58.000 vezes mais rápido. Talvez outra comparação possa ajudar a estabelecer uma noção melhor da aceleração. Em média, uma pessoa caminha a uma velocidade de 5 km/h. Um jato hipersônico viaja por volta de mil vezes mais rápido, a 6.100 km/h, pouco mais de cinco vezes a velocidade do som. Imagine multiplicar isso por 58.000.

Para onde foi toda essa velocidade e potência de processamento? São duas as respostas: *viabilidade* e *usabilidade*. Em termos de aplicativos, podemos fazer cada vez mais. Podemos fazê-lo de maneiras cada vez mais fáceis, não apenas quanto à programação, mas principalmente quanto à experiência do usuário. Os vídeos, por exemplo, são muito exigentes com relação à computação. O mesmo acontece com os sistemas operacionais. A IA é possível hoje também porque temos o poder computacional necessário para executar esse tipo de *software*.

Graças a esse crescimento surpreendente das capacidades de armazenamento e processamento a custos cada vez mais acessíveis, bilhões de pessoas estão conectadas atualmente. Passam muitas horas *online* diariamente. Conforme o Statista.com, por exemplo, "em 2018, o tempo médio gasto usando a Internet [no Reino Unido] era de 25,3 horas por semana. Isso representou um aumento de 15,4 horas em relação a 2005"[4]. Isso está longe de ser incomum e a pandemia fez uma diferença ainda mais significativa. Voltarei a esse ponto no Capítulo 2, mas outro motivo pelo qual a IA é possível hoje é porque estarmos passando cada vez mais tempo em contextos que são digitais e amigáveis para a IA.

Mais memória, potência, velocidade, ambientes e interações digitais geraram dados em quantidades imensas. Todos nós já vimos diagramas com curvas exponenciais indicando valores que nem sequer podemos imaginar. Segundo a empresa de inteligência de mercado IDC[5], o ano de 2018 viu a humanidade atingir 18 *zettabytes* de dados (criados, capturados ou replicados). O crescimento espantoso dos dados não mostra sinais de desaceleração; aparentemente, chegará a 175 *zettabytes* em 2025. Isso é difícil de entender em termos de quantidade, mas duas consequências merecem um momento de reflexão[6]. Primeiro, a velocidade e a memória das tecnologias digitais não estão crescendo no mesmo ritmo que o universo de dados. Portanto, estamos passando rapidamente de uma cultura de gravação para uma de exclusão; a questão não é mais o que salvar, mas o que excluir para abrir espaço para novos dados. Em segundo lugar, a maioria dos dados disponíveis foi criada a partir dos anos 90 (mesmo se incluirmos todas as palavras pronunciadas, escritas ou impressas na história da

[4] Disponível em: https://www.statista.com/statistics/300201/hours-of-internet-use-per-week-per-person-in-the-uk/. Acesso em: 20 set. 2024.
[5] Consulte a discussão, disponível em: https://www.seagate.com/gb/en/our-story/data-age-2025/. Acesso em: 20 set. 2024.
[6] Discuto ambas em Floridi (2014a).

humanidade, bem como todas as bibliotecas ou arquivos que já existiram). Basta olhar qualquer um desses diagramas *online* que ilustram a explosão de dados: o lado surpreendente não está apenas à direita, para onde vai a seta do crescimento, mas também à esquerda, onde ela começa. Isso foi há apenas alguns anos. Como todos os dados que temos foram criados pela geração atual, eles também estão envelhecendo em termos de suporte e tecnologias obsoletas. Portanto, sua curadoria será uma questão cada vez mais urgente.

Mais poder computacional e mais dados possibilitaram a mudança da lógica para a estatística. As redes neurais, que antes eram apenas teoricamente interessantes[7], tornaram-se ferramentas comuns no aprendizado de máquina (ML). A IA antiga era principalmente simbólica e podia ser interpretada como um ramo da lógica matemática, mas a nova IA é principalmente conexionista e pode ser interpretada como um ramo da estatística. A inferência e a estatística, e não mais a dedução lógica, são as principais armas secretas da IA.

Potência e velocidade da computação, tamanho da memória, volumes de dados, efeitos dos algoritmos e das ferramentas estatísticas, bem como o número de interações *online* têm crescido de forma incrivelmente rápida. Isso também se deve ao fato de que (aqui, a conexão causal ocorre em ambos os sentidos) o número de dispositivos digitais interagindo entre si já é várias vezes maior do que a população. Portanto, a maioria da comunicação agora é feita de máquina para máquina, sem envolvimento humano. Temos robôs computadorizados em Marte controlados remotamente a partir da Terra. O comandante Christopher "Kit" Draper teria achado essas coisas absolutamente incríveis.

Todas as tendências anteriores continuarão crescendo, incansavelmente, no futuro próximo. Elas mudaram a forma como aprendemos, nos divertimos, trabalhamos, amamos, odiamos, escolhemos e decidimos produzir, vender, comprar, consumir, anunciar, nos divertir, cuidar e ser cuidado, socializar, nos comunicar e assim por diante. Parece impossível localizar qualquer aspecto da vida que não tenha sido afetado pela revolução digital. No último meio século, aproximadamente, a realidade se tornou cada vez mais digital. Ela é composta de zeros e uns sendo governada por *software* e dados, em vez de *hardware* e átomos. Cada vez mais as pessoas vivem *onlife*, (Floridi, 2014b), tanto *online* quanto *offline*, e na infosfera, tanto digital quanto analógica.

[7] Discuti algumas delas em um livro bem anterior, do final dos anos 90 (Floridi, 1999).

Essa revolução digital também afeta a forma como conceituamos e entendemos as realidades, as quais são cada vez mais interpretadas em termos computacionais e digitais. Basta pensar na "antiga" analogia entre o seu DNA e o seu "código", que hoje consideramos óbvia. A revolução impulsionou o desenvolvimento da IA à medida que compartilhamos experiências *onlife* e ambientes na infosfera com agentes artificiais (AA), sejam eles algoritmos, *bots* ou robôs. Para entender o que a IA pode representar (argumentarei ser uma nova forma de agência, não de inteligência) é preciso falar mais sobre o impacto da própria revolução digital. Essa é a tarefa do restante deste capítulo. É somente compreendendo a trajetória conceitual de suas implicações que podemos ter a perspectiva correta sobre a natureza da IA (Capítulo 2), seus prováveis desenvolvimentos (Capítulo 3) e seus desafios éticos (Parte Dois).

1.2. **O *cleaving power* do digital: copiando e colando a modernidade**

As tecnologias, ciências, práticas, produtos e serviços digitais, em resumo, *o digital* como fenômeno geral, estão transformando profundamente a realidade. Isso é óbvio e incontroverso. As verdadeiras questões são: *por que*, *como* e *e daí*, especialmente no que diz respeito à IA. Em cada caso, a resposta está longe de ser trivial e está certamente aberta a debates. Para explicar as respostas que considero mais convincentes e apresentar no próximo capítulo uma interpretação da IA como um crescente *reservatório de agência inteligente*, deixe-me começar *in medias res*, ou seja, a partir do "como". Assim, será mais fácil retroceder para entender o "por quê" e, em seguida, avançar para lidar com o "e daí" antes de vincular as respostas ao surgimento da IA.

O digital "copia e cola" as realidades, tanto ontológica quanto epistemologicamente. Com isso, quero dizer que ele conecta, desconecta ou reconecta características do mundo (ontologia) e, portanto, as suposições correspondentes sobre elas (epistemologia) que pensávamos ser imutáveis. Ele separa e funde os "átomos" da experiência e cultura "modernas", por assim dizer. Ele remodela o leito do rio, para usar uma metáfora wittgensteiniana. Alguns exemplos claros podem nos ajudar a entender melhor esse ponto.

Consideremos primeiro um dos casos mais significativos de conexão. A identidade própria e os dados pessoais nem sempre foram colados de forma tão indistinta como são hoje quando falamos da *identidade pessoal*

dos *titulares dos dados*. As contagens do censo são muito antigas (Alterman, 1969). A invenção da fotografia teve um enorme impacto sobre a privacidade (Warren; Brandeis, 1890). Os governos europeus tornaram obrigatório viajar com passaporte durante a Primeira Guerra Mundial, por motivos de migração e segurança, ampliando assim o controle do Estado sobre os meios de mobilidade (Torpey, 2000). Mas foi apenas o digital, com seu imenso poder de registrar, monitorar, compartilhar e processar quantidades ilimitadas de dados sobre Alice, que uniu quem é Alice, seu eu e perfil individuais, às informações pessoais sobre ela. A privacidade tornou-se uma questão urgente também, talvez a principal, devido a essa conexão. Hoje, pelo menos na legislação da UE, a proteção de dados é discutida em termos de dignidade humana (Floridi, 2016c) e identidade pessoal (Floridi, 2005a, 2006, 2015b) e os cidadãos são descritos como *titulares dos dados*.

O próximo exemplo diz respeito à *localização* e *presença* e à desconexão entre elas. Em um mundo digital, é óbvio que alguém pode estar fisicamente localizado em um lugar (por exemplo, uma cafeteria) e interativamente presente em outro (por exemplo, uma página no Facebook). No entanto, todas as gerações passadas que viveram em um mundo exclusivamente analógico conceberam e vivenciaram a localização e a presença como dois lados inseparáveis da mesma situação humana: estar situado no espaço e no tempo, aqui e agora. A ação à distância e a telepresença pertenciam a mundos mágicos ou à ficção científica. Hoje, essa desconexão simplesmente reflete a experiência comum em qualquer sociedade da informação (Floridi, 2005b). Somos a primeira geração para a qual a pergunta "onde você está?" não é apenas retórica. É claro que a desconexão não cortou todos os vínculos. A geolocalização só funciona se a telepresença de Alice puder ser monitorada. E a telepresença de Alice só é possível se ela estiver localizada em um ambiente fisicamente conectado. Mas as duas estão agora completamente distintas e, de fato, a desconexão entre elas desvalorizou ligeiramente a localização em favor da presença. Porque se tudo de que Alice precisa e com que se preocupa é estar digitalmente presente e interativa em um determinado canto da infosfera, não importa em que parte do mundo ela esteja localizada analogicamente, seja em casa, em um trem ou em seu escritório. É por isso que bancos, livrarias, bibliotecas e lojas de varejo são locais *de presença* em busca de um reposicionamento de *local*. Quando uma loja abre

uma cafeteria, ela está tentando unir a presença e a localização dos clientes, uma conexão que foi rompida pela experiência digital.

Em seguida, consideremos a desconexão entre a *lei* e a *territorialidade*. Durante séculos, aproximadamente desde a Paz de Westfália, em 1648, a geografia política tem proporcionado à jurisprudência uma resposta fácil para a questão da abrangência de uma decisão: ela deve ser aplicada até as fronteiras nacionais dentro das quais a autoridade legal opera. Essa conexão poderia ser resumida como "meu lugar, minhas regras; seu lugar, suas regras". Hoje isso pode parecer óbvio, mas levou muito tempo e imenso sofrimento para chegar a uma abordagem tão simples. Ela ainda é perfeitamente adequada hoje, se a pessoa estiver operando apenas em um espaço físico analógico. No entanto, a Internet não é um espaço físico. O problema da territorialidade surge de um desalinhamento ontológico entre o espaço normativo da lei, o espaço físico da geografia e o espaço lógico do digital. Essa é uma "geometria" nova e variável que ainda estamos aprendendo a administrar. Por exemplo, a desconexão entre a lei e a territorialidade tornou-se óbvia e problemática durante o debate sobre o chamado *direito de ser esquecido* (Floridi, 2015a). Os mecanismos de pesquisa operam em um espaço lógico *online* de nós, *links*, protocolos, recursos, serviços, URLs, interfaces e assim por diante. Isso significa que tudo está a apenas um clique de distância. Portanto, é difícil implementar o direito ao esquecimento solicitando ao Google que remova os *links* para as informações pessoais de alguém de sua versão .com nos Estados Unidos devido a uma decisão tomada pelo Tribunal de Justiça da União Europeia (TJUE), mesmo que essa decisão pareça inútil, a menos que os *links* sejam removidos de todas as versões do mecanismo de pesquisa.

Observe que esse desalinhamento entre os espaços causa problemas e oferece soluções. A não territorialidade de obras digitais faz maravilhas para a livre circulação de informações. Na China, por exemplo, o governo precisa fazer esforços constantes e persistentes para controlar as informações *online*. Na mesma linha, o Regulamento Geral sobre a Proteção de Dados (RGPD) deve ser admirado pela capacidade de explorar a "conexão" entre a identidade pessoal e as informações pessoais para contornar a "desconexão" entre a lei e a territorialidade. Isso é feito fundamentando a proteção de dados pessoais em termos da primeira (a quem eles estão "vinculados", o que agora é fundamental) em vez da segunda (onde eles estão sendo processados, o que não é mais relevante).

Por fim, aqui está uma conexão que acaba sendo, mais precisamente, uma reconexão. Em seu livro *A Terceira Onda*, de 1980, Alvin Toffler cunhou o termo "prossumidor" para se referir à indefinição e à fusão do papel de produtores e consumidores (Toffler, 1980). Toffler atribuiu essa tendência a um mercado altamente saturado e ao volume de produtos padronizados, o que levou a um processo de customização em massa que, por sua vez, levou ao envolvimento crescente dos consumidores como produtores de seus próprios produtos personalizados. A ideia tinha sido antecipada em 1972 por Marshall McLuhan e Barrington Nevitt, que atribuíram o fenômeno às tecnologias baseadas na eletricidade. Posteriormente, passou a se referir ao consumo de informações produzidas pela mesma população de produtores, como no YouTube. Sem saber desses precedentes, introduzi a palavra "*produmer*" ("produmidor") para captar o mesmo fenômeno quase vinte anos depois de Toffler[8]. Mas em todos esses casos, o que está em jogo não é uma *nova* conexão. Mais precisamente, é uma reconexão.

Durante a maior parte da nossa história (aproximadamente 90 por cento dela; consulte Lee e Daly (1999)), vivemos em sociedades de caçadores e coletores em busca de alimentos para sobreviver. Durante esse período, produtores e consumidores normalmente se sobrepunham. Os prossumidores que caçavam animais selvagens e colhiam plantas silvestres eram, em outras palavras, a normalidade e não a exceção. Foi apenas a partir do desenvolvimento das sociedades agrárias, há cerca de dez mil anos, que assistimos a uma completa (e, depois de algum tempo, culturalmente óbvia) separação entre produtores e consumidores. Mas em alguns cantos da infosfera, essa desconexão está sendo reconectada. No Instagram, no TikTok ou no Clubhouse, por exemplo, consumimos o que produzimos. Portanto, pode-se enfatizar que, em alguns casos, esse parêntese está chegando ao fim e que os prossumidores estão de volta, reconectados pelo digital. Logo, é perfeitamente coerente que o comportamento humano *online* tenha sido comparado e estudado em termos de modelos de forrageamento desde a década de 1990 (Pirolli; Card, 1995, 1999; Pirolli, 2007).

O leitor pode listar facilmente mais casos de conexão, desconexão e reconexão. Pense, por exemplo, na diferença entre *realidade virtual* (desco-

[8] Em Floridi (1999); veja também Floridi (2004, 2014b). Eu deveria ter me informado melhor e usado o termo "*prossumer*" de Toffler.

nexão) e *realidade aumentada* (conexão); na desconexão comum entre *uso* e *propriedade* na economia compartilhada; na reconexão entre *autenticidade* e *memória* graças à cadeia de *blockchain*; ou no debate atual sobre uma renda básica universal, que é um caso de desconexão entre *salário* e *trabalho*. Mas é hora de passar da questão do "como" para a questão do "por quê". Por que o digital tem esse *cleaving power*[9, 10] para conectar, desconectar e reconectar o mundo e nossa compreensão dele? Por que outras inovações tecnológicas parecem não ter um impacto semelhante? A resposta, suponho, reside na combinação de dois fatores.

Por um lado, o digital é uma *tecnologia de terceira ordem* (Floridi, 2014a). Não se trata apenas de uma tecnologia entre nós e a natureza, como um machado (primeira ordem), ou uma tecnologia entre nós e outra tecnologia, como um motor (segunda ordem). É mais uma tecnologia entre uma tecnologia e outra tecnologia, como um sistema computadorizado controlando um robô pintando um carro (terceira ordem). Devido ao poder de processamento autônomo do digital, podemos nem estar na execução, muito menos dentro da execução.

Por outro lado, o digital não está apenas melhorando ou aumentando a realidade. Ele transforma radicalmente a realidade porque cria novos ambientes que passamos a habitar e novas formas de agência com as quais passamos a interagir. Não existe um termo para essa profunda forma de transformação. No passado (Floridi, 2010b), usei a expressão *reontologização* para me referir a um tipo muito radical de reengenharia. Esse tipo de reengenharia não apenas projeta, constrói ou estrutura um sistema novamente (por exemplo, uma empresa, uma máquina ou algum artefato), mas transforma fundamentalmente a natureza intrínseca do próprio sistema, ou seja, sua ontologia. Nesse sentido, as nanotecnologias e as biotecnologias não estão apenas reestruturando, mas

[9] Escolhi "*cleaving*" como um termo particularmente adequado para me referir ao poder de desconexão/reconexão do digital porque "*to cleave*" [clivar] tem dois significados: (a) "dividir ou cortar algo", especialmente ao longo de uma linha ou granulação natural; e (b) "grudar rapidamente ou aderir fortemente a" algo. Isso pode parecer contraditório, mas se deve ao fato de que "*to cleave*" é o resultado da fusão, em uma única grafia, e, portanto, em um significado duplo, de duas palavras separadas: (a) vem do inglês antigo *cleofan*, relacionado ao alemão *klieben* (cortar); enquanto (b) vem do inglês antigo *clifian*, relacionado ao alemão *kleben* (grudar ou agarrar). Têm raízes protoindo-europeias muito diferentes; consulte http://www.etymonline. com/index.php?term=cleave.

[10] N. da T.: O português infelizmente não tem os mesmos dois sentidos para o verbo clivar. A etimologia da palavra esclarece que ela vem "do neerlandês *klieven*, 'fender; rachar', pelo francês *cliver*, 'idem'" (Infopédia, Dicionários Porto Editora). Como é importante que a palavra utilizada nesse contexto possa transmitir ao mesmo tempo as ideias de conectar e desconectar, a opção adotada foi deixar o termo em inglês.

reontologizando nosso mundo. Resumindo, ao *reontologizar a modernidade* o digital também está *reepistemologizando a mentalidade moderna* (ou seja, muitos de nossos antigos conceitos e ideias).

Juntos, todos esses fatores sugerem que o digital deve seu *cleaving power* ao fato de ser uma tecnologia de terceira ordem reontologizante e reepistemologizante. É por isso que faz o que faz e que nenhuma outra tecnologia chegou perto de ter um efeito semelhante.

1.3. **Novas formas de agência**

Se tudo isso estiver mais ou menos correto, talvez ajude a entender alguns fenômenos atuais relativos à transformação da morfologia da agência na era digital e, consequentemente, as *formas de agência* que apenas parecem não estar relacionadas. A transformação delas depende do *cleaving power* do digital, mas sua interpretação pode ser decorrente de um mal-entendido implícito sobre esse *cleaving power* e suas consequências crescentes, profundas e duradouras. Estou me referindo à *agência política* como democracia direta e à *agência artificial* como IA. Em cada caso, a reontologização da agência ainda não foi seguida por uma reepistemologização adequada de sua interpretação. Ou, dito de forma menos precisa, mas talvez mais intuitiva: o digital mudou a natureza da agência, mas ainda estamos interpretando o resultado dessas mudanças em termos da mentalidade moderna e isso está gerando alguns mal-entendidos profundos.

O tipo de agência a que estou me referindo não é o mais discutido na filosofia ou na psicologia, envolvendo estados mentais, intencionalidade e outras características tipicamente associadas aos seres humanos. A agência discutida neste livro é a que é comumente encontrada na ciência da computação (Russell; Norvig, 2018) e na literatura sobre sistemas multiagentes (Weiss, 2013; Wooldridge, 2009). Ela é mais minimalista e exige que um sistema satisfaça apenas três condições básicas. Ele pode:

a) receber e usar dados do ambiente, por meio de sensores ou outras formas de entrada de dados;
b) realizar ações com base nos dados de entrada, de forma autônoma, para atingir metas por meio de atuadores ou outras formas de saída, e
c) melhorar seu desempenho, aprendendo a partir de suas interações.

Tal agente pode ser artificial (por exemplo, um *bot*), biológico (por exemplo, um cachorro), social (por exemplo, uma empresa ou um governo) ou híbrido. A seguir, falarei brevemente sobre agência política e democracia direta porque este livro se concentrará apenas na agência artificial, não na agência sociopolítica moldada e amparada por tecnologias digitais.

Nos debates atuais sobre democracia direta, às vezes somos induzidos ao erro de pensar que o digital *deveria* (observe a abordagem normativa em vez de descritiva) reconectar *soberania* e *governança*. Soberania é o poder político que pode ser legitimamente delegado, enquanto governança é o poder político que é legitimamente delegado, de forma temporária, condicional e responsável e que pode ser retirado de forma igualmente legítima (Floridi, 2016e). A democracia *representativa* é comumente, embora erroneamente, vista como um meio-termo devido a restrições práticas de comunicação. A verdadeira democracia seria *direta*, com base na participação de todos os cidadãos em questões políticas de forma universal, constante e sem mediação. Infelizmente, como se sabe, somos muitos. Portanto, a delegação (entendida como intermediação) do poder político é um mal necessário, ainda que menor (Mill, 1861, p. 69). É o mito da cidade-estado e, especialmente, de Atenas.

Durante séculos, esse meio-termo em favor de uma democracia representativa pareceu inevitável, isto é, até a chegada do digital. De acordo com algumas pessoas, isso agora promete desintermediar a democracia moderna e conectar (ou reconectar, se você acredita nos míticos bons e velhos tempos) a soberania à governança para oferecer um novo tipo de democracia. Seria um tipo de ágora digital que poderia finalmente permitir o envolvimento regular e direto de todos os cidadãos interessados. É a mesma promessa feita pelo instrumento do referendo, especialmente quando ele é vinculativo e não consultivo. Em ambos os casos, os eleitores são questionados diretamente sobre o que deve ser feito. A única tarefa deixada para as classes política, administrativa e técnica seria a de implementar as decisões do povo. Os políticos seriam funcionários públicos *delegados* (não *representantes*) em um sentido muito literal. No entanto, isso é um erro porque o plano sempre foi a democracia indireta. Falando de maneira mais trivial, a desconexão é uma característica muito mais do que um *bug*. Um regime democrático é caracterizado, acima de tudo, não por alguns *procedimentos* ou *valores* (embora estes também possam ser características), mas por uma *separação* clara

e nítida, ou seja, uma desconexão, entre aqueles que detêm o poder político (soberania) e aqueles a quem ele é confiado (governança). Todos os cidadãos em idade de votar têm poder político e o delegam legitimamente por meio do voto. Esse poder é, então, confiado a outras pessoas que exercem o mandato, governando de forma transparente e responsável pelo tempo em que estiverem legitimamente habilitadas. Simplificando, um regime democrático não é apenas uma forma de exercer e gerenciar o poder de algumas maneiras (procedimentos) e/ou de acordo com alguns valores. É, antes de tudo, uma forma de *estruturar* o poder. Aqueles que detêm o poder político não o exercem; eles o dão àqueles que o exercem, mas não o detêm. A fusão dos dois lados leva a formas frágeis de autocracia, ou governo da multidão.

Sob essa perspectiva, a democracia representativa não é um meio-termo. É na verdade a melhor forma de democracia. E usar o digital para conectar (ou, como já observei, de um ponto de vista mais mítico, reconectar) a soberania à governança seria um grande erro. Brexit, Trump, Liga Norte e outros desastres populistas causados pela "tirania da maioria" (Adams, 1787) são provas suficientes. Precisamos considerar a melhor forma de aproveitar a desconexão planejada e representativa entre soberania e governança, e não a melhor forma de eliminá-la. Portanto, o consenso é o problema. Entretanto, esse não é o assunto deste livro. O que quis proporcionar com a análise anterior foi uma amostra do tipo de considerações unificadoras sobre formas de agência que vinculam o impacto do digital na política à forma como pensamos e avaliamos a IA, como veremos na próxima seção. Voltemos agora à agência artificial.

1.4. **IA: uma área de pesquisa em busca de uma definição**

Algumas pessoas (talvez muitas) parecem acreditar que a IA consiste em conectar a agência artificial e o comportamento inteligente a novos artefatos. Isso é um mal-entendido. Como explicarei mais detalhadamente no próximo capítulo, na verdade é o contrário: a revolução digital tornou a IA não apenas possível, mas também cada vez mais útil. Isso foi feito *desconectando* a capacidade de resolver um problema ou concluir uma tarefa de maneira eficaz de qualquer necessidade de ser inteligente para fazê-lo. Só quando essa desconexão for alcançada é que

a IA terá sucesso. Então, a queixa habitual conhecida como "efeito IA"[11] – assim que a IA consegue executar uma tarefa específica, como tradução automática ou reconhecimento de voz, o alvo move-se e essa tarefa deixa de ser definida como inteligente se for executada por IA – é, na verdade, um reconhecimento correto do processo preciso em questão. A IA só executa uma tarefa com êxito se conseguir desconectar a sua realização de qualquer necessidade de ser inteligente ao fazê-lo; portanto, se a IA for bem-sucedida, então a desconexão ocorreu e, de fato, a tarefa demonstrou ser dissociável da inteligência que parecia ser necessária (por exemplo, em um ser humano) para levar ao sucesso.

Isso é menos surpreendente do que parece e no próximo capítulo veremos que é perfeitamente consistente com a definição clássica (e provavelmente ainda uma das melhores) de IA fornecida por McCarthy, Minsky, Rochester e Shannon em "*Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*" [Proposta para o Projeto de Pesquisa de Verão de Dartmouth sobre Inteligência Artificial], o documento fundador e posterior evento que estabeleceu o novo campo da IA em 1955. Vou apenas citá-la aqui e adiar sua discussão para o próximo capítulo:

> para o presente propósito, o problema da inteligência artificial é considerado o de fazer com que uma máquina se comporte de maneira que seria chamada de inteligente se um ser humano estivesse se comportando dessa forma.[12]

As consequências de compreender a IA como um divórcio entre agência e inteligência são profundas. O mesmo acontece com os desafios éticos que esse divórcio gera e a segunda parte do livro será dedicada à análise deles. Mas, para concluir este capítulo introdutório, ainda é preciso dar uma resposta final à pergunta "e daí" que mencionei no início. Essa é a tarefa da próxima seção.

[11] Disponível em: https://en.wikipedia.org/wiki/AI_effect. Acesso em: 20 set. 2024.
[12] Versão *online* disponível em: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html. Acesso em: 20 set. 2024; veja também a reedição em McCarthy *et al.* (2006).

1.5. **Conclusão: ética, governança e *design***

Supondo que as respostas anteriores para as perguntas "por quê" e "como" sejam aceitáveis, que diferença faz quando entendemos o poder do digital em termos de copiar e colar o mundo (juntamente com nossa conceituação dele) de maneiras sem precedentes? Uma analogia pode ajudar a dar a resposta. Se alguém possui apenas uma única pedra e absolutamente nada mais, nem mesmo outra pedra para colocar ao lado dela, então também não há nada mais que se possa fazer além de apreciar a pedra, olhando para ela ou brincando com ela. Mas se a pedra for cortada em duas, surgem várias possibilidades de combinação. Duas pedras proporcionam mais *affordance* e menos restrições do que uma única pedra. Várias pedras, ainda mais. O *design* se torna possível.

Copiar e colar os blocos ontológicos e conceituais da modernidade, por assim dizer, é exatamente o que o digital faz de melhor quando se trata de seu impacto em nossa cultura e filosofia. Aproveitar *affordance* e restrições para resolver alguns problemas é chamado de design. Portanto, a resposta agora deve estar clara: o *cleaving power* do digital diminui enormemente as restrições da realidade e aumenta sua *affordance*. Ao fazer isso, ele transforma o *design* – mais ou menos entendido como a arte de resolver um problema por meio da criação de um artefato que aproveita restrições e *affordance* para satisfazer alguns requisitos, tendo em vista um objetivo – na atividade inovadora que define nossa era.

Nossa jornada agora está completa. Cada época inovou suas culturas, sociedades e ambientes com base em pelo menos três elementos principais: *descoberta*, *invenção* e *design*. Esses três tipos de inovação estão intimamente ligados, embora a inovação tenha sido muitas vezes distorcida como um banco de três pernas, no qual uma perna é mais longa e, portanto, mais avançada do que as outras. O período pós-renascentista e o início do período moderno podem ser qualificados como a era das descobertas, especialmente geográficas. A modernidade tardia ainda é uma era de descobertas, mas, com suas inovações industriais e mecânicas, talvez seja mais do que nunca uma era de invenções. E, é claro, todas as eras também foram eras de *design* (até porque descobertas e invenções exigem maneiras engenhosas de conectar e dar forma a novas e velhas realidades). Mas se estou certo no que argumentei até agora, então é a nossa era que é essencialmente, mais do que qualquer outra, *a era do design*.

Como o digital está diminuindo as restrições e aumentando a *affordance* à nossa disposição, ele nos oferece uma liberdade imensa e crescente para arrumar e organizar o mundo de várias maneiras a fim de resolver uma variedade de problemas novos e antigos. Obviamente, qualquer *design* requer um projeto. E, no nosso caso, o que ainda nos falta é um *projeto humano* para a era digital. Mas não devemos permitir que o *cleaving power* do digital molde o mundo sem um plano. Devemos fazer de tudo para decidir a direção na qual queremos explorá-lo, a fim de garantir que as sociedades da informação que estamos construindo graças a ele sejam abertas, tolerantes, equitativas, justas e favoráveis ao meio ambiente, bem como à dignidade e à prosperidade humanas. A consequência mais importante do *cleaving power* do digital deveria ser um design melhor do mundo. E isso diz respeito à configuração da IA como uma nova forma de agência, como veremos no Capítulo 2.