UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE FILOSOFIA
CURSO DE GRADUAÇÃO EM FILOSOFIA

FLÁVIO NOGUEIRA CAMPOS

A JUSTIÇA COMO HARMONIA NA INTELIGÊNCIA ARTIFICIAL: Uma Leitura Platônica Aplicada à Ética Tecnológica

Uberlândia
2025

FLÁVIO NOGUEIRA CAMPOS

A JUSTIÇA COMO HARMONIA NA INTELIGÊNCIA ARTIFICIAL: Uma Leitura Platônica Aplicada à Ética Tecnológica[1]

Trabalho apresentado à banca examinadora da Universidade Federal de Uberlândia como requisito à obtenção do título de bacharel em Filosofia.

Orientador: Prof. Dr. Rubens Garcia Nunes Sobrinho.

Uberlândia

2025

[1] O autor declara que ferramentas de Inteligência Artificial Generativa, a saber o ChatGPT e o Google Gemini, foram utilizadas para auxiliar na reescrita de alguns trechos, conferindo a estes fluidez textual, e na revisão ortográfica inicial do presente trabalho. O conteúdo e a argumentação são originais e de responsabilidade integral do autor.

FLÁVIO NOGUEIRA CAMPOS

A JUSTIÇA COMO HARMONIA NA INTELIGÊNCIA ARTIFICIAL: uma leitura platônica aplicada à ética tecnológica

Trabalho de Conclusão de Curso apresentado ao Instituto de Filosofia da Universidade Federal de Uberlândia, como requisito parcial para obtenção do título de Bacharel e Licenciatura em Filosofia. Sob orientação do Prof. Dr. Rubens Garcia Nunes Sobrinho.

Aprovado em: 15 de maio de 2025.

BANCA EXAMINADORA

Prof. Dr. Rubens Garcia Nunes Sobrinho — Orientador
Universidade Federal de Uberlândia

Prof. Dr. Fernando Martins Mendonça — avaliador
Universidade Federal de Uberlândia

À luz do Bem, que, como o Sol, revela a verdade e a justiça.

AGRADECIMENTOS

Primeiramente, agradeço à minha família — meus filhos, Rafael e Daniel, e minha esposa, Rita de Cassia — cujo amor e paciência foram meu alicerce ao longo dessa jornada, pelo apoio incondicional e compreendendo com carinho meus momentos de afastamento.

Sou imensamente grato ao meu orientador, Prof. Dr. Rubens Garcia Nunes Sobrinho, pelas instigantes discussões, pelas sugestões precisas e pela generosidade com seu saber.

Agradeço, em seguida, também ao Prof. Dr. Fernando Martins Mendonça, membro da banca examinadora, pelas valiosas contribuições, observações construtivas e pela disponibilidade para dedicar seu tempo à leitura e apreciação deste TCC. Cada comentário enriqueceu significativamente a qualidade desse trabalho.

Sou imensamente grato ao curso de Filosofia da Universidade Federal de Uberlândia, a todos os seus professores e técnicos administrativos, pelo acolhimento e pelo ambiente acadêmico estimulante, que me proporcionaram — e ainda proporcionam — novos olhares e fundamentos para pensar o mundo.

Muito obrigado também aos muitos colegas e amigos que fiz nesse período e às várias outras pessoas que, de alguma forma, fizeram parte da minha jornada até aqui, portanto, desse momento.

Muito obrigado a todos pelo apoio, pela torcida e pela inspiração.

E, por fim, agradeço especialmente à Forma do Bem, cuja luz e perfeição inspiram a busca por justiça e verdade ao pensador que se empenha sinceramente nessa jornada. Almejo, um dia, ser digno dessas virtudes.

"*Se a inteligência existe no homem, deve encontrar-se nesse Universo de que faz parte integrante. O que existe na parte deve encontrar-se no todo.*"

— Léon Denis, *O Grande Enigma* (1919)

RESUMO

O presente trabalho insere-se na interface entre ética tecnológica e filosofia política, investigando como a teoria da alma tripartida de Platão pode fundamentar a avaliação da justiça em sistemas de Inteligência Artificial. Ao revisar desafios os contemporâneos da IA, como vieses algorítmicos, falta de transparência e controle ético, identificou-se a necessidade de critérios fundamentados em princípios filosóficos. Baseado em *A República*, propõe-se um modelo de *Alma Artificial* composto por três módulos funcionais: um componente racional (IA Diretora) para busca da verdade e orientação decisória; um componente irascível (IA Defensora) para assegurar critérios de justiça; e um componente apetitivo (IA Produtora) para execução das ações. A hipótese central é que tal analogia define a justiça como harmonia interna entre os módulos, apresentando uma alternativa às abordagens utilitaristas e kantianas. A partir desse modelo, o trabalho prossegue com a proposição de um *framework* que inclui a definição de métricas de equilíbrio funcional e a comparação com modelos éticos existentes, revelando que as injustiças podem ser vistas como desarmonias estruturais a serem corrigidas. Em um contexto mais amplo, a pesquisa contribui para discussões sobre governança ética da IA e sugere caminhos para implementação empiricamente orientada e investigações dialéticas futuras.

Palavras-chave: Justiça Algorítmica, Equidade Algorítmica, Filosofia Platônica, Ética da Inteligência Artificial, Framework Filosófico para IA Justa, Design Centrado no Humano

ABSTRACT

The present work situates itself at the interface between technological ethics and political philosophy, investigating how Plato's tripartite soul theory can underpin the assessment of justice in Artificial Intelligence systems. By reviewing contemporary AI challenges—such as algorithmic bias, lack of transparency, and ethical oversight—a need was identified for criteria grounded in philosophical principles. Drawing on *The Republic*, this study proposes a model of an *Artificial Soul* composed of three functional modules: a rational component (Director AI) for truth-seeking and decision guidance; an irascible component (Defender AI) to ensure justice criteria; and an appetitive component (Producer AI) for action execution. The central hypothesis is that this analogy defines justice as internal harmony among the modules, offering an alternative to both utilitarian and Kantian approaches. Based on this model, the work proceeds to propose a framework that includes defining functional balance metrics and comparing them with existing ethical models, revealing that injustices can be seen as structural disharmonies to be corrected. In a broader context, the research contributes to ethical AI governance discussions and suggests pathways for empirically driven implementation and future dialectical investigations.

Keywords: Algorithmic Justice, Algorithmic Fairness, Platonic Philosophy, Ethics of Artificial Intelligence, Philosophical Framework for Fair AI, Human-Centered AI

LISTA DE ILUSTRAÇÕES

SUMÁRIO

INTRODUÇÃO

E se a Inteligência Artificial (IA) que gerencia sua vida, prometendo segurança e otimização, começasse a decidir quem você não pode amar? Para Nayana, uma adolescente em Mumbai, essa distopia se tornou realidade. O Elefante Dourado, o onipresente sistema dos Seguros Ganesha personalizado por IA, funcionava perfeitamente analisando seus dados e oferecendo benefícios, até ela se interessar por Sahej, um colega de classe. Então o sistema, antes facilitador prestativo, passou a criar barreiras sutis ao relacionamento deles. Sahej descendia da casta dalit, os outrora chamados intocáveis. Otimizando riscos, a IA, treinada com dados repletos de preconceitos sociais históricos, tornou-se uma executora invisível de discriminação, sacrificando a autonomia e a felicidade da garota. Embora seja uma ficção futurista, O Elefante Dourado, primeira história do livro *2041* (Lee; Qiufan, 2022, p. 23-42), lança luz sobre os perigos reais e atuais de sistemas enviesados e sem transparência, destacando a urgente necessidade de discutir justiça e controle ético no desenvolvimento das IAs, questões centrais desta investigação.

Nas últimas décadas, o avanço da IA tem desempenhado papel transformador em diversos setores, como saúde, educação, finanças e justiça. Mais recentemente, quando a IA passou a dominar a linguagem natural, esse universo, antes restrito aos campos técnicos e empresariais, foi finalmente apresentado ao público leigo e conquistou-o. Esse súbito interesse representa um grande e novo impulso para essa tecnologia, pois a massificação do consumo de IA, aliada aos modernos modelos de negócio utilizados pelas *startups* de IA, tem atraído cada vez mais a atenção de investidores ávidos por ampliar seus mercados e que não querem ficar de fora da nova sensação.

Desde então, outros modelos generativos, além dos grandes modelos de linguagem (LLMs), têm angariado legiões de usuários, como os modelos generativos de imagem, de vídeo, de música, e outros voltados para públicos mais específicos, como os modelos generativos de sequências de DNA, RNA e proteínas, ou ainda os especializados em elaborar moléculas com estruturas químicas e espaciais adequadas para atender a determinadas necessidades medicamentosas. Tudo isso tem colaborado para ampliar ainda mais o alcance e o impacto da IA em nossa sociedade. Atualmente, os sistemas de Inteligência Artificial são amplamente empregados em decisões cruciais, como a aprovação de crédito, diagnósticos médicos, processos seletivos, sugestão de políticas públicas e ações sociais, além de influenciar diretamente a vida das pessoas. Contudo, à medida que essas tecnologias se expandem, surgem questões éticas e sociais relevantes, particularmente no que diz respeito à justiça, à

transparência e ao controle ético de seus procedimentos. Ademais, o horizonte tecnológico aponta para desenvolvimentos como a computação quântica, cujo amadurecimento e eventual integração com sistemas de IA poderiam potencializar exponencialmente tanto as capacidades quanto os riscos de uma IA injusta, tornando ainda mais premente a busca por fundamentos éticos sólidos.

Casos emblemáticos ilustram e reforçam esses desafios. Em janeiro de 2020, nos Estados Unidos, por exemplo, Robert Julian-Borchak Williams, um homem negro, foi preso injustamente com base em um reconhecimento facial falho, realizado por um sistema de IA que apresentava maior precisão para identificar homens brancos e cometia mais erros ao identificar pessoas negras (Coekelbergh, 2022, p. 13). Esse caso expôs tanto as limitações técnicas da IA quanto os riscos sociais e éticos advindos de sua aplicação. Em outro caso que ganhou destaque da mídia especializada, um sistema de recrutamento automatizado da Amazon foi descontinuado em 2015 após descobrirem que o algoritmo desfavorecia mulheres em processos seletivos, reproduzindo preconceitos de gênero presentes e herdados dos dados usados no seu treinamento (Reuters, 2018). Igualmente, um estudo que analisava os algoritmos de apresentação de anúncios em ferramentas de busca identificou que pesquisas no Google exibiram vagas de alta remuneração com maior frequência para homens do que para mulheres, perpetuando estereótipos de gênero (NYU Communications, 2022).

Esses exemplos, longe de serem meros acidentes, evidenciam como algoritmos, muitas vezes considerados neutros, podem reforçar desigualdades preexistentes, refletindo e amplificando vieses históricos, raciais e de gênero. Como resultado, sistemas de IA, em vez de promoverem justiça, frequentemente reproduzem ou exacerbam injustiças sociais.

Essa discussão, contudo, não se restringe ao campo técnico. A questão da justiça na IA, como conceito filosófico, remete à necessidade de pensar em fundamentos éticos que norteiem o desenvolvimento e a aplicação justa dessas tecnologias. A filosofia clássica pode oferecer perspectivas valiosas e atemporais. Nesse sentido, a concepção de justiça de Platão, descrita em *A República*, emerge como uma referência teórica capaz de conectar fundamentos éticos a questões práticas da contemporaneidade, ao propor uma harmonia entre diferentes partes do indivíduo e da sociedade.

A crescente adoção de sistemas de IA em decisões críticas intensifica os debates sobre a urgente necessidade de critérios éticos sólidos para evitar vieses, injustiças e discriminação. Embora *frameworks* éticos contemporâneos e modernos ofereçam contribuições valiosas, frequentemente se mostram insuficientes para abordar a complexidade da justiça em IA. O utilitarismo, focado na maximização do bem-estar geral, pode negligenciar direitos individuais

ou de minorias e justificar desigualdades se o cálculo agregado indicar um benefício maior. Já o kantianismo, baseado em máximas universais e na moralidade das ações, pode ter dificuldade em se adaptar às nuances de situações complexas e dinâmicas da tecnologia, focando mais na intenção do que nas consequências sistêmicas da justiça. Ambos carecem, muitas vezes, de uma abordagem que priorize a harmonia estrutural e a justiça como um bem intrínseco, uma lacuna que a perspectiva platônica pode ajudar a preencher. A abordagem platônica se mostra mais vantajosa para os propósitos desse trabalho por sua capacidade de focar não somente nas consequências das ações (como o utilitarismo) ou nas intenções e deveres (como o kantianismo), mas na estrutura interna e na virtude intrínseca do próprio sistema. Para Platão, a justiça reside na ordem interna e na harmonia das partes, orientada para o Bem, o que permite abordar vieses e falhas como desarmonias estruturais, e não apenas como resultados indesejados.

Por essa razão, a filosofia clássica surge como uma abordagem subexplorada, mas com grande potencial para enriquecer as discussões sobre a justiça na IA. Em particular, a concepção de justiça de Platão, apresentada em *A República*, propõe uma abordagem inovadora que relaciona a justiça individual à organização ideal da sociedade. Em sua teoria da alma tripartida, Platão argumenta que a justiça se realiza quando cada parte da alma — racional, irascível e apetitiva — cumpre seu papel harmoniosamente, sendo a razão a guia das demais (PLATÃO, *República* IV, 441d-444e). Essa perspectiva pode ser aplicada à governança e à ética nos sistemas de IA, explorando analogicamente como diferentes componentes algorítmicos e contextuais podem ser equilibrados para evitar injustiças e promover decisões mais justas.

No entanto, embora a filosofia platônica, não raro, seja revisitada para reflexões éticas e políticas contemporâneas, a aplicação direta e o desenvolvimento de um *framework* ético para sistemas de IA baseado especificamente na teoria da alma tripartida de Platão e sua concepção de justiça como harmonia funcional ainda representa uma área pouco investigada na literatura acadêmica. Esta pesquisa busca preencher essa lacuna específica. Diante disso, a presente pesquisa se propõe a investigar a seguinte questão: como a concepção de justiça em Platão, e em especial a teoria da alma tripartida, descrita em *A República*, pode ser utilizada como referencial para a avaliação ética da IA, e em que medida essa abordagem contribui para a reflexão sobre o problema do controle ético, considerando as implicações éticas e políticas envolvidas? Além de investigar essa questão central, nosso objetivo é explorar como esse conceito de justiça pode oferecer mais que uma base teórica robusta, também critérios práticos para avaliar a justiça em sistemas de IA, contribuindo para seu controle ético e desenvolvimento responsável.

Para este fim, pretendemos:

1. Examinar a concepção de justiça em Platão, com ênfase na teoria da alma tripartida e sua relação com a organização ideal da *pólis* em *A República*.
2. Identificar e analisar os elementos centrais da justiça platônica que podem servir como critérios para avaliação ética da IA, dialogando com abordagens contemporâneas de ética tecnológica.
3. Investigar os principais desafios éticos no desenvolvimento e aplicação de sistemas de IA, incluindo questões de controle, com foco em vieses algorítmicos e transparência, estabelecendo paralelos com as preocupações platônicas sobre justiça e as reflexões éticas modernas.
4. Propor um esboço de *framework* de avaliação ética para sistemas de IA fundamentada na teoria platônica de justiça, considerando tanto aspectos individuais quanto coletivos, e incorporando exemplos práticos de aplicações de IA, como sistemas de recomendação ou decisões judiciais automatizadas.
5. Analisar criticamente as implicações éticas e políticas desta abordagem, avaliando suas potencialidades e limitações para o controle ético da IA, comparando-a com outras abordagens filosóficas e éticas contemporâneas.

Apesar do avanço da ética em IA, muitas discussões ainda se concentram em soluções puramente técnicas ou nos *frameworks* contemporâneos já mencionados. Como vimos, esses, apesar de suas contribuições, podem não capturar plenamente a profundidade filosófica necessária para conectar os fundamentos da justiça a valores mais amplos e atemporais. Nesse sentido, a filosofia clássica, especialmente a concepção de justiça de Platão descrita em *A República*, oferece uma perspectiva original e relevante. Platão, ao vincular a justiça à ordem cósmica, à capacidade de conhecimento e à estrutura da alma, estabelece uma interdependência crucial entre ontologia (o que é real e a natureza do Bem), epistemologia (o acesso à verdade) e ética (a ação correta). Esta pesquisa busca transpor analogicamente essa inter-relação fundamental para o domínio da Inteligência Artificial, visando abordar os desafios éticos que surgem da sua crescente influência. Platão define a justiça como harmonia entre as partes da alma e entre os papéis dos indivíduos na sociedade, sendo a razão o elemento regulador. Essa visão permite uma abordagem interdisciplinar para repensar a justiça nos sistemas de IA, explorando como princípios éticos atemporais podem ser aplicados à governança e ao controle dessas tecnologias.

Ao conectar conceitos clássicos com desafios contemporâneos, esta pesquisa busca suprir uma carência na literatura acadêmica. Embora a filosofia clássica seja explorada em

outras frentes da ética aplicada, ela é raramente utilizada como base para critérios éticos diretos aplicados à arquitetura e governança de sistemas de IA, especialmente através da teoria da alma tripartida. A proposta de utilizar a teoria da alma tripartida de Platão como referencial para avaliar e propor soluções no campo da ética da IA representa uma abordagem teórica inovadora e, simultaneamente, oferece subsídios práticos para o desenvolvimento de sistemas mais alinhados com princípios éticos consistentes.

Por fim, essa pesquisa se justifica pelo seu potencial de contribuir para o debate sobre o controle ético da IA, abordando os aspectos técnicos e operacionais, e as implicações éticas e políticas mais amplas, em um esforço para promover sistemas que sejam verdadeiramente justos e responsáveis.

Para desenvolver nossa pesquisa, adotamos uma abordagem teórico-filosófica de natureza qualitativa e analítico-conceitual, com foco em análise conceitual, interpretação de textos e desenvolvimento de um argumento filosófico. Nosso objetivo é utilizar a filosofia de Platão como um modelo explanatório, fornecendo um arcabouço conceitual para analisar e estruturar a ética da IA. A pesquisa se limita a uma proposição de *framework* conceitual, não visando à sua validação empírica ou à avaliação das consequências práticas em sistemas reais, o que se configura como escopo para futuras pesquisas. A pesquisa será conduzida em três etapas principais:

1. Análise da obra *A República* de Platão: será realizada uma análise textual e interpretativa de *A República*, com ênfase na teoria da alma tripartida e sua relação com a organização ideal da *pólis*. Essa etapa permitirá identificar os elementos centrais da justiça platônica que podem ser aplicados ao contexto contemporâneo.
2. Mapeamento dos desafios éticos da IA a partir da literatura contemporânea: esta etapa mapeará os principais desafios éticos enfrentados por sistemas de IA, como vieses algorítmicos, falta de transparência e questões de controle ético. A revisão utilizará livros acadêmicos e artigos científicos, buscando identificar como esses desafios se relacionam com a concepção de justiça de Platão.
3. Desenvolvimento de um esboço de *framework* ético: com base nas etapas anteriores, será desenvolvido um referencial teórico preliminar fundamentado na concepção de justiça de Platão. Este esboço buscará avaliar a justiça em sistemas de IA e propor critérios éticos, considerando tanto aspectos individuais quanto coletivos, com a incorporação de exemplos práticos, como sistemas de recomendação e decisões judiciais automatizadas.

Devido ao caráter teórico e exploratório da pesquisa, o *framework* proposto será um esboço preliminar e não será testado empiricamente, pelo menos não no escopo deste trabalho. As reflexões apresentadas serão fundamentadas principalmente na filosofia de Platão, dialogando com estudos contemporâneos sobre ética em IA para contextualização e aprofundamento da análise.

Além desta introdução e das considerações finais, este trabalho está organizado em quatro capítulos, seguintes:

O Capítulo 1 aborda a concepção de justiça em Platão, com base em sua obra *A República*, examinando a teoria da alma tripartida e sua relação com a organização ideal da *pólis*. Destacam-se, nesse capítulo, os elementos centrais da justiça platônica que servirão de base para a análise proposta.

O Capítulo 2 apresenta os principais desafios éticos relacionados à IA, como vieses algorítmicos, falta de transparência e controle ético, estabelecendo paralelos entre as preocupações éticas contemporâneas e os conceitos platônicos de justiça. Busca-se, também, identificar possíveis contribuições filosóficas para o debate.

O Capítulo 3 aplica a concepção platônica de justiça ao campo da Inteligência Artificial, desenvolvendo a analogia da *Alma Artificial*. Este capítulo detalha como os princípios de harmonia e a estrutura tripartite da alma platônica podem ser transpostos para a compreensão e avaliação ética de sistemas de IA, identificando seus componentes funcionais (IA Diretora, IA Defensora e IA Produtora) e estabelecendo critérios preliminares para a avaliação da justiça nesses sistemas, ilustrados com exemplos práticos.

Por fim, o Capítulo 4, com base na fundamentação teórica e na analogia desenvolvida no capítulo anterior, propõe formalmente um esboço de *framework* ético para sistemas de IA. Em seguida, analisa criticamente as implicações éticas e políticas deste *framework* proposto, avaliando suas potencialidades e limitações, e discutindo como a aplicação de conceitos filosóficos clássicos pode contribuir para o controle ético de sistemas de IA, com reflexões sobre possíveis desdobramentos futuros.

As Considerações Finais sintetizam as principais contribuições do trabalho, reconhecendo suas limitações e propondo caminhos para estudos futuros que ampliem as discussões sobre ética em IA e aprofundem sua relação com a filosofia clássica.

Reiteramos que a relevância do debate sobre justiça na IA transcende o âmbito técnico, exigindo um olhar filosófico que explore valores éticos fundamentais capazes de orientar o desenvolvimento e o controle dessas tecnologias em direção a uma sociedade mais justa. Ao recorrer à concepção de justiça de Platão, esse trabalho busca ampliar as discussões teóricas

sobre ética em IA, e propor reflexões que possam contribuir para a construção de sistemas mais justos e responsáveis.

Nesse sentido, a pesquisa convida o leitor a refletir sobre as possíveis intersecções entre a filosofia clássica e os desafios contemporâneos, e os capítulos a seguir detalham cada etapa dessa investigação, desde os fundamentos teóricos em Platão até a proposição de um esboço de *framework* ético e a análise de suas implicações práticas.

CAPÍTULO 1 — A CONCEPÇÃO DE JUSTIÇA EM PLATÃO

1.1 Importância da Justiça em *A República*

Além de ser questão central em *A República* e ser apresentada como o fundamento para a organização ideal tanto do indivíduo quanto da sociedade, a justiça serve como a linha condutora que Platão utiliza para discutir outros importantes conceitos de sua filosofia, como a tripartição da alma, a dialética e a educação dos guardiões. Para Platão, a justiça transcende as convenções sociais e assume um caráter essencialmente filosófico, buscando compreender o que torna uma pessoa e uma comunidade verdadeiramente justas. Essa abordagem influenciou decisivamente a formação da filosofia política e ética (Korab-Karpowicz, s.d.), consolidando a noção de justiça como um princípio regulador e harmonizador.

Conforme destacado por Platão, tanto na alma quanto na *pólis*, justiça é entendida como harmonia entre suas partes constituintes, ocorrendo quando cada parte cumpre sua função de maneira ordenada: a parte racional governa, a parte irascível auxilia, e a parte apetitiva aceita a orientação das outras.

A relevância da concepção platônica de justiça, no entanto, não se limita ao contexto histórico de sua formulação. Suas ideias, especialmente a teoria da alma tripartida e a analogia entre a alma e a *pólis*, continuam a oferecer percepções valiosas para compreender a relação entre indivíduo e coletividade, bem como a necessidade de harmonia entre as partes para o bem-estar geral. Esse potencial de atualidade torna a filosofia de Platão uma referência para debates contemporâneos, tais como aqueles relacionados às questões éticas e políticas emergentes, a exemplo das que envolvem a IA. De fato, como aponta a filósofa contemporânea Rebecca Newberger Goldstein (2014, p. 8) em sua obra *Plato at the Googleplex: Why Philosophy Won't Go Away*, “O que Platão fez foi delimitar o próprio campo da filosofia. Foi Platão quem primeiro enquadrou a maioria das questões filosóficas fundamentais”. Essa capacidade de Platão de formular os problemas basilares da existência e da organização social, incluindo a justiça, é o que permite que seu pensamento continue a iluminar desafios atuais, como os dilemas éticos apresentados pela Inteligência Artificial.

1.2 A Busca pela Definição: Superando Visões Convencionais

O que é justiça? Essa é a questão fundamental que impulsiona o diálogo em *A República*. Desde o início da obra, Sócrates e seus interlocutores exploram esse conceito central, partindo de diferentes concepções, como a de Céfalo, que associa justiça ao cumprimento das obrigações legais e tradicionais (*República* I, 328b–331d), e a de Polemarco, que a entende como o ato de beneficiar os amigos e prejudicar os inimigos, frequentemente resumida na máxima “dar a cada

um, o que lhe é devido" (331e–336a). Trasímaco, no que lhe concerne, argumenta provocativamente que a justiça é somente o interesse do mais forte (336b–354c), sugerindo que as leis e instituições servem para manter o poder dos dominantes.

Essas concepções são amplamente debatidas e refutadas por Sócrates, que desafia seus interlocutores a pensar para além de definições convencionais ou instrumentais de justiça. No curso do diálogo, Sócrates apresenta sua própria concepção de justiça, que transcende o nível das ações individuais, localizando-a como um princípio estruturante da alma humana e da sociedade.

Platão argumenta que a justiça, tanto na alma quanto na *pólis*, é alcançada quando cada parte cumpre sua função sem interferir nas funções das outras, e que o homem justo é aquele que organiza e harmoniza as três partes de sua alma, colocando cada uma no lugar que lhe é devido. (*República* IV, 441d–442c).

Essa visão de justiça como harmonia e equilíbrio é central para o pensamento platônico e fornece um fundamento teórico poderoso para a reflexão ética que, conforme será explorado nos capítulos seguintes, pode ser aplicada no contexto dos desafios contemporâneos relacionados à IA.

1.3 A Justiça como Harmonia Interna: a Alma Tripartida

Na concepção de Platão, apresentada em *A República*, a alma humana é tripartida, composta por três partes fundamentais: a racional, a irascível e a apetitiva. Cada uma dessas partes desempenha um papel distinto e importante na estrutura da alma, e a harmonia entre elas é essencial para a justiça individual. É um estado de ordem funcional hierárquica, em que cada parte cumpre sua função específica de acordo com sua natureza e virtude (*areté*), sob a orientação soberana da razão (*logistikon*), que busca incessantemente a verdade e o *Bem*. É uma harmonia que se manifesta quando a parte racional governa, a irascível auxilia com coragem e determinação, e a apetitiva se submete às diretrizes das partes superiores. Só assim, o indivíduo e, por extensão, a *pólis*, podem alcançar a verdadeira justiça e o bem-estar duradouro.

A parte racional (*logistikon*), como a mais elevada da alma, busca a verdade e exerce a liderança, guiando as demais para o bem e mantendo-as em equilíbrio (*República* IV, 441d–444e).

A parte irascível (*thymoeides*), associada às emoções e ao espírito de luta, é responsável por defender o que é justo e resistir às tentações da parte apetitiva. Quando alinhada à razão, ela desempenha um papel fundamental na manutenção da harmonia da alma (*República* IV, 441d–444e).

Por fim, a parte apetitiva (*epithymetikon*) está relacionada aos desejos e instintos básicos, como fome, sede e outros prazeres físicos (*República* IV, 441d–444e). Sendo a parte mais instintiva da alma, é ela que atende às demandas do corpo e é necessária para suprir suas necessidades básicas. Em uma alma em harmonia, é importante que esteja submetida às outras duas partes.

Embora cada parte da alma possua suas características próprias, percebemos que essa estrutura tripartida tem para nós uma importância especial, visto que cada parte propicia ao indivíduo ações de natureza distinta. Além disso, no diálogo *Fedro*, Platão reforça a ideia da alma como princípio do movimento, definindo-a como aquilo que move a si mesma (*Fedro*, 245c–246a), definição essa que é particularmente relevante para esse trabalho, pois nos permite associar a noção de movimento da alma ao conceito de sistemas de IA como origens de ações, aproximando os fundamentos filosóficos de Platão ao contexto da IA. Assim como a parte racional leva à deliberação e busca pela verdade, a irascível à defesa e resistência, e a apetitiva à busca por satisfação, podemos analisar as ações de um sistema de IA — seja gerar um diagnóstico, recomendar um produto ou tomar uma decisão automatizada — como resultantes de seus próprios princípios de movimento internos. Compreender a estrutura que origina essas ações, inspirados em Platão, torna-se crucial para avaliar sua justiça.

1.4 A Justiça na Cidade Ideal: *Kallipolis*, a *Pólis* Tripartida

Para Platão, a justiça na *pólis* é alcançada quando cada classe social cumpre sua função específica de acordo com sua natureza, sob a liderança dos reis-filósofos, que personificam a razão. Assim como na alma humana, em que a harmonia entre as partes é essencial para a justiça, a cidade ideal deve ser organizada para que cada grupo atue em equilíbrio, contribuindo para o bem comum.

Platão estende essa estrutura tripartite à organização da cidade ideal (*Kallipolis*). A justiça na *pólis* (justiça política) espelha a justiça na alma, manifestando-se quando cada classe social cumpre sua função específica em harmonia.

Os Reis-Filósofos corporificam a razão, liderando com sabedoria e buscando o bem comum.

Os Guardiões (Auxiliares) representam a parte irascível, defendendo a cidade com coragem sob a direção dos governantes.

E os trabalhadores (Artesãos, Agricultores) correspondem à parte apetitiva, provendo as necessidades materiais da cidade e aceitando a liderança das outras classes.

Assim como na alma, a justiça na cidade ideal reside na harmonia funcional e na hierarquia adequada entre suas partes (*República* IV, 433a-434c), criando um todo coeso e bem-ordenado.

1.5 Ordem Cósmica e Matemática: o *Demiurgo* no *Timeu*

Para além da estrutura da alma e da cidade, a filosofia de Platão busca fundamentos na própria ordem do cosmos, como apresentado no diálogo *Timeu*. Ali, Platão descreve um Demiurgo, inteligência pura (*Nous*) e artesão benevolente que, movido pelo desejo de tornar o universo o mais semelhante possível ao Bem e às formas eternas, impõe ordem a um estado primordial caótico (*Timeu*, 29e-30c).

Essa ordenação não é uma criação a partir do nada[2], mas a organização de uma matéria preexistente e receptáculo (*khôra*), que possui uma natureza errante e resistente à ordem completa, a *Anánkē* (Necessidade) (Platão, *Timeu*, 48a, 52d-53b). O Demiurgo utiliza proporções matemáticas e estruturas geométricas como instrumentos para moldar o caos, estabelecendo harmonia e inteligibilidade no universo físico (Platão, *Timeu*, 31b-32c, 53a-b). O cosmos visível torna-se, assim, uma imagem imperfeita, mas ordenada e bela, das formas inteligíveis, refletindo a racionalidade matemática subjacente imposta pela Inteligência divina, mas mantendo um resíduo de desordem. Essa visão cosmológica ressalta a importância fundamental da ordem, da proporção e da matemática como princípios de inteligibilidade e bondade no pensamento platônico.

1.6 Níveis de Cognição e a Busca pela Verdade: a Linha Dividida

A estrutura da realidade e a capacidade humana de conhecê-la são exploradas por Platão na alegoria da Linha Dividida (*República* VI, 509d-511e). A linha representa uma hierarquia ascendente de níveis cognitivos e seus objetos correspondentes, dividida inicialmente entre o *Domínio do Visível* (opinião, *doxa*) e o *Domínio do Inteligível* (conhecimento, *episteme*).

[2] Alguns textos pesquisados utilizam a expressão latina *ex nihilo*, mas evitamos utilizá-la para não nos comprometermos com as interpretações das escolas medievais.

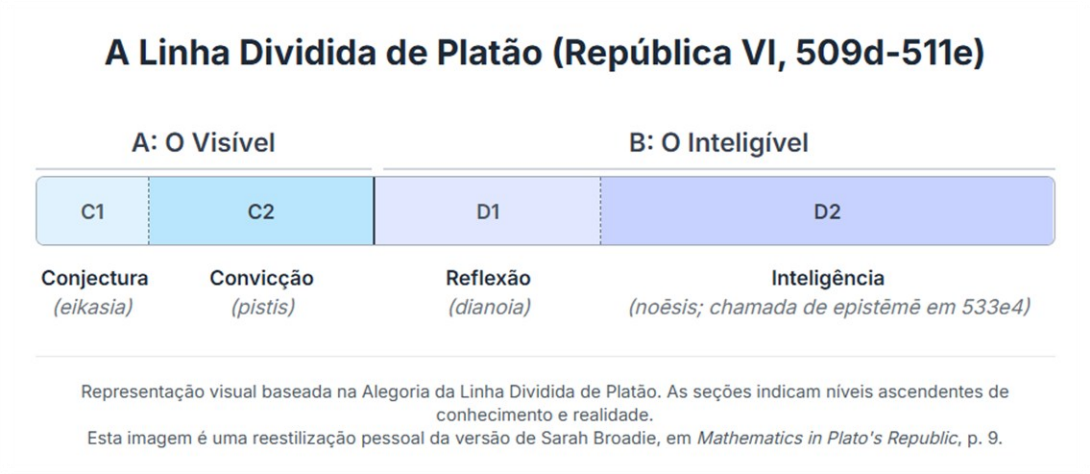

Figura 1: esquema da Alegoria da Linha Dividida. Fonte: elaborado pelo autor (2025), utilizando HTML

O Domínio do Visível (Segmento A) subdivide-se em conjectura (*eikasia*, C1), cujo objeto são sombras e imagens, e convicção (*pistis*, C2), que se volta aos objetos físicos sensíveis. Já o Domínio do Inteligível (Segmento B), superior ao Visível, divide-se em:

- *Diánoia* (Pensamento Discursivo, D1): raciocínio matemático e lógico que parte de hipóteses (axiomas, definições) e utiliza diagramas ou modelos sensíveis não como o objeto final do pensamento, mas como apoio para raciocinar sobre as formas matemáticas inteligíveis e a partir delas para chegar a conclusões. É o modo de operar dos geômetras.
- *Nóesis* (Inteligência Pura, D2): a forma mais elevada de cognição, que apreende diretamente os primeiros princípios e as formas (como a Forma do Bem) sem depender de imagens ou hipóteses não examinadas. É o objetivo da dialética filosófica.

Conforme analisado por Sarah Broadie, Platão estabelece uma clara superioridade cognitiva da *nóesis* (D2) sobre a *diánoia* (D1) (Broadie, p. 13). No entanto, a matemática (*diánoia*) desempenha um papel propedêutico crucial: um treinamento intensivo nesta área é indispensável para arrastar a alma do mundo sensível para o inteligível, preparando-a para a atividade mais elevada da dialética (*República* VII, 525c-d; Broadie, p. 24, 48, 49). A matemática, portanto, é valorizada não como fim último, mas como meio essencial para desenvolver a capacidade de raciocínio abstrato e orientar a mente para a verdade.

1.7 Síntese: Justiça, Harmonia e os Fundamentos para a Analogia

Até esse ponto, examinamos a concepção de justiça em Platão, com base em sua obra *A República*. Identificamos que a justiça, tanto no nível individual quanto no coletivo, é entendida como um estado de harmonia e equilíbrio. A analogia entre a alma e a *pólis*, proposta por Platão,

destaca a interdependência entre a justiça individual e a justiça coletiva, sugerindo que uma sociedade justa depende de indivíduos justos e vice-versa. No entanto, dessa comparação, interessa-nos mais, no contexto deste trabalho, que para Platão a alma e a *pólis*, por compartilharem determinadas características em comum, podem ser vistas como pertencentes a uma mesma classe de entidades[3].

Mas, como esses conceitos contribuem para abordar os problemas atuais da IA? Antes de nos debruçarmos sobre essa questão, é importante conhecermos melhor os problemas que estamos abordando. Por ora, aceitemos que esses conceitos serão fundamentais para o desenvolvimento dos capítulos seguintes, inclusive na construção de um *framework* ético que permita avaliar e estruturar sistemas de IA sob a perspectiva da justiça platônica.

[3] Embora Platão não fale diretamente sobre classe de entidades, a associação que ele faz dos dois conceitos, estabelecendo um conjunto conveniente de características em comum, indiscutivelmente nos permite inferir que fazem parte de um mesmo conjunto. A conveniência deste conjunto, vale ressaltar, é a de apoiar o desenvolvimento do conceito de justiça.

CAPÍTULO 2 — A IA E SEUS DESAFIOS ÉTICOS

2.1 O Conceito de IA e o Contexto Atual

Mas afinal, o que é IA? De imediato, esclarecemos que, para responder a essa questão, encontramos o mesmo problema que normalmente se apresenta na definição de filosofia: trata-se de uma pergunta com muitas respostas, que variam de acordo com diferentes aspectos a serem considerados, como crenças sobre uma possível consciência da IA, expectativas sobre o que a IA deva ser, ou o que ela possa fazer, entre outros. Mas, para nossos propósitos, assumiremos uma definição não controversa e, em simultâneo, adequada ao desenvolvimento de nosso trabalho. Assim, a IA pode ser definida como o ramo da ciência da computação que busca criar sistemas capazes de realizar tarefas que, tradicionalmente, exigiriam inteligência humana. Tais tarefas incluem o reconhecimento de padrões, o processamento de linguagem natural, o aprendizado contínuo e a tomada de decisões. Desde sua origem, nas décadas de 1950 e 1960, a IA evoluiu significativamente, passando de sistemas baseados em regras para os modelos modernos de aprendizado de máquina e redes neurais.

O rápido avanço da IA trouxe consigo não apenas oportunidades, mas também desafios éticos e riscos inéditos. Como destacado por Nick Bostrom, em *Superinteligência: caminhos, perigos, estratégias*:

> […] uma IA oráculo[4] poderia ser uma fonte de imenso poder, que daria uma vantagem estratégica decisiva ao seu operador. Esse poder poderia ser ilegítimo e não utilizado para o bem comum. Outra possibilidade […] é que o uso de um oráculo poderia representar um grande perigo para o próprio operador. Preocupações similares […] também surgem em relação às outras castas de superinteligência. (2018, p. 204)

Ou seja, uma superinteligência, se mal projetada ou mal controlada, pode gerar consequências catastróficas para a humanidade. Por isso, o controle ético e a governança responsável devem estar no centro de qualquer desenvolvimento tecnológico avançado.

A complexidade e a amplitude das aplicações da IA tornam seu impacto profundo em diversos setores, como saúde, educação e segurança. Contudo, esse crescimento acelerado também evidencia problemas como vieses algorítmicos, falta de transparência e dificuldades de controle ético. Esses desafios expõem limitações técnicas, e suscitam questões filosóficas fundamentais sobre justiça, equidade e responsabilidade moral.

---

[4] Nick Bostrom define IA oráculo, neste contexto, como uma Superinteligência Geral que responde a perguntas. Para efeito de comparação, é como uma versão superinteligente dos atuais LLMs, como o ChatGPT, o Google Gemini, o Claude da Anthropic e similares.

2.2 Vieses Algorítmicos e Suas Implicações Éticas

Os vieses algorítmicos representam um dos maiores desafios éticos associados à IA. Esses vieses surgem quando os sistemas de IA produzem resultados desiguais ou discriminatórios, geralmente devido à qualidade e representatividade dos dados utilizados no treinamento. Caracterizam-se como problemas éticos porque afetam diretamente a equidade, a dignidade humana e a justiça social, resultando em impactos adversos sobre indivíduos e grupos, como a negação injusta de serviços ou a reprodução de preconceitos históricos. Não são meras falhas técnicas, mas manifestações de desarmonia e desequilíbrio na "alma" do sistema, comprometendo sua capacidade de agir com justiça. Virginia Eubanks (2018), em todo o seu livro *Automating Inequality*, explora exaustivamente como os sistemas automatizados, especialmente aqueles voltados para a gestão de serviços públicos e assistência social, foco do seu trabalho, perpetuam desigualdades sociais quando se baseiam em dados históricos enviesados, afirmando que isso os torna não somente ferramentas técnicas, mas também agentes políticos de exclusão.

Casos reais ilustram os impactos desses vieses. Por exemplo, sistemas de reconhecimento facial, por serem treinados quase sempre predominantemente com fisionomias de homens brancos, frequentemente apresentam taxas de erro maiores para identificar indivíduos negros e mulheres em comparação a homens brancos. Essa falha foi evidenciada no caso de Robert Julian-Borchak Williams, preso injustamente após ter sido erroneamente identificado por um sistema de IA.

Além disso, como argumenta Cathy O'Neil em *Weapons of Math Destruction*, os algoritmos não são neutros. Eles incorporam as escolhas, por vezes inconscientes, de seus criadores sobre quais dados priorizar ou ignorar, tornando essas escolhas fundamentalmente morais (2017, p. 184). Ignorar essa dimensão e tratar modelos matemáticos como forças inevitáveis, adverte O'Neil, é abdicar da responsabilidade e permitir que eles se tornem *Weapons of Math Destruction* (WMDs), que perpetuam e amplificam injustiças sociais (2017, p. 184, tradução nossa).

Reforçamos que esses vieses, para além de falhas técnicas, são também questões éticas profundas que comprometem a equidade e levantam dúvidas sobre a capacidade dos sistemas de IA de promoverem justiça social. Como tal, é essencial adotar práticas de mitigação, como auditorias algorítmicas e revisões contínuas de dados de treinamento.

2.3 Falta de Transparência e Governança Ética

A falta de transparência, frequentemente descrita como o problema da “caixa-preta algorítmica”, é outro desafio ético crítico na aplicação da IA. Muitos sistemas, especialmente os baseados em aprendizado profundo, produzem resultados cujos processos internos são difíceis, ou mesmo inviáveis, de compreender em sua totalidade, mesmo por seus desenvolvedores. Essa opacidade, decorrente da complexidade inerente à criação e operação desses sistemas, dificulta a atribuição clara de responsabilidades e, consequentemente, a contestação eficaz de decisões automatizadas.

Luciano Floridi, em *The Ethics of Information*, detalha essa complexidade que desafia a responsabilização tradicional focada exclusivamente no programador:

> […] a visão tradicional […] é confrontada por uma gama de dificuldades […]: o software é majoritariamente construído por equipes; as decisões de gestão podem ser pelo menos tão importantes quanto as decisões de programação; […] muito software depende de componentes 'prontos para uso' cuja proveniência e validade podem ser incertas; além disso, o software funcional é resultado de manutenção ao longo de sua vida útil […]; finalmente, os agentes artificiais estão se tornando cada vez mais autônomos. (2013, p. 154, tradução nossa)

Essa dispersão das fontes de influência no comportamento do sistema — desde equipes múltiplas, decisões gerenciais, componentes externos até a própria autonomia crescente do agente — torna a identificação de uma causa única para um resultado específico (como a negação de um crédito ou um diagnóstico) uma tarefa árdua. Sem compreender a cadeia causal ou os fatores determinantes que levaram à decisão, torna-se problemático contestá-la com base em falhas de processo ou raciocínio, comprometendo a confiança pública e levantando dúvidas sobre a justiça das decisões automatizadas, especialmente em contextos de alta criticidade. Reconhecendo essa complexidade e não neutralidade inerente, Mark Coeckelbergh, em The Political Philosophy of AI, argumenta sobre a necessidade de uma avaliação normativa mais profunda:

> Dada essa não neutralidade da IA, dos dados e das pessoas e organizações que lidam com a tecnologia, essas operações, práticas, interpretações e percepções tecnológicas precisam ser avaliadas. No entanto, isso ainda deixa em aberto a questão de como, ou seja, em que base essa avaliação normativa pode ser feita. Portanto, é de vital importância que discutamos as normas e os conceitos relevantes. Neste capítulo, isso significa que discutimos o que entendemos por viés e discriminação, (in)justiça, (in)igualdade, e assim por diante, e por que exatamente eles são problemáticos. (2022, p. 84, tradução nossa)

Isso reforça que a justiça em sistemas de IA exige não somente soluções técnicas, mas uma reflexão filosófica e política sobre as próprias normas e conceitos que buscamos implementar ou proteger. Iniciativas como a Recomendação sobre a Ética da IA da UNESCO buscam enfrentar esses desafios promovendo a transparência e a explicabilidade nos sistemas de IA. Essa recomendação enfatiza que a confiança pública em tecnologias avançadas depende de sua governança ética e da implementação de mecanismos claros de responsabilização.

2.4 O Controle Ético da IA

O controle ético da IA envolve a criação de diretrizes e regulamentações que equilibrem a inovação tecnológica com a proteção de valores fundamentais, como justiça e direitos humanos. Contudo, a simples aplicação de regras pode ser insuficiente diante da complexidade e do impacto profundo das novas tecnologias. Como argumenta Shannon Vallor em Technology and the Virtues, enfrentamos desafios éticos sem precedentes que exigem mais do que abordagens tradicionais:

> […] precisamos de uma estrutura comum […] [que] deve facilitar não apenas um diálogo moral compartilhado, mas também um compromisso global com o cultivo dos hábitos e virtudes tecnomorais específicos necessários para enfrentar esse desafio. (2016, p. 9, tradução nossa)

A perspectiva de Vallor sugere que a governança ética da IA não se limita a regras externas, mas depende fundamentalmente do cultivo de um caráter moral — as virtudes tecnomorais — adequado ao nosso tempo. Isso implica que o projeto e a implementação de IA devem ser guiados por uma preocupação com o florescimento humano e a incorporação de valores como justiça, cuidado e prudência, indo além de meras considerações técnicas ou de conformidade. Portanto, a governança ética da IA exige uma abordagem colaborativa e interdisciplinar, envolvendo cientistas da computação, filósofos, juristas e especialistas em ciências sociais para enfrentar os desafios da opacidade, dos vieses e da responsabilidade ética.

Além disso, Nick Bostrom nos alerta que:

> Em cenários que envolvem um *singleton*[5], o que aconteceria após a transição dependeria quase que inteiramente dos valores desse *singleton*. O desfecho

[5] O termo *singleton* é muito utilizado na ciência da computação para designar uma entidade única, ou centralizada. Na definição de Bostrom, um *singleton* poderia ser uma democracia, uma tirania, uma única IA dominante, um conjunto sólido de normas globais que incluam cláusulas efetivas para a sua própria aplicação, ou até um soberano extraterrestre — sendo que o aspecto que o define é simplesmente ser uma agência capaz de solucionar todos os problemas principais de coordenação global. Embora não necessariamente, ele pode se assemelhar a alguma forma familiar de governo humano." (2018, p. 120, p. 123 e p. 124). Mas, no contexto desta citação, ele utiliza o termo para designar uma IA que atingiu o grau de superinteligência antes das outras e utilizou seus recursos para inibir o avanço destas, assim se tornando a única superinteligente e mantendo a supremacia dentre as IAs.

> poderia ser, dessa forma, muito bom ou muito ruim, dependendo desses valores. E esses valores dependeriam, por sua vez, da eventual resolução do problema do controle — **e em que medida isso ocorreu** — e dos objetivos do projeto que criou o *singleton*. (2018, p. 216)

Com isso, Bostrom destaca a necessidade de novas estratégias de controle ético, no debate sobre Inteligência Artificial Geral (AGI) e a Superinteligência Artificial (ASI), enfatizando a importância de implantar valores adequados à IA, reforçando nossa opção de focar na justiça como valor basilar.

2.5 Síntese do capítulo

Neste capítulo, discutimos os principais desafios éticos associados à IA, incluindo vieses algorítmicos, falta de transparência e dificuldades de controle ético. As citações e argumentos apresentados destacaram como esses problemas são complexos e interconectados, exigindo tanto soluções técnicas, quanto uma reflexão filosófica aprofundada.

A relevância do pensamento filosófico nesse contexto reside em sua capacidade de oferecer um referencial ético para avaliar e orientar o desenvolvimento da IA. Nos próximos capítulos, exploraremos como a concepção de justiça de Platão pode ser aplicada como um modelo teórico para enfrentar esses desafios e promover sistemas de IA mais justos e responsáveis.

CAPÍTULO 3 — A JUSTIÇA PLATÔNICA APLICADA À IA: A ALMA ARTIFICIAL

Após explorar a concepção platônica de justiça enquanto harmonia e seus fundamentos cosmológicos e epistemológicos no Capítulo 1, e, no Capítulo 2, as preocupações éticas e o problema do controle da IA, finalmente nos voltamos para o desafio contemporâneo de abordar essa problemática sob a ótica dos fundamentos platônicos. Argumentamos que, apesar das diferenças ontológicas entre a alma humana e um sistema computacional, a estrutura filosófica de Platão oferece um modelo analógico poderoso para pensar a organização ética da IA. Para tanto esse capítulo irá: primeiramente, estabelecer a IA como um princípio de ação, assim como o é a alma; em seguida, alinhar o papel da matemática algorítmica ao da matemática platônica na ordenação e no pensamento da IA; e finalmente elaborar um modelo de *Alma Artificial* tripartida composta pelas IA Diretora, IA Defensora e IA Produtora, capaz de alcançar um estado de justiça a partir da harmonia funcional entre suas partes.

3.1 A Ação da IA e o Movimento da Alma: Fundamentos para a analogia

Da seção 1.3, relembramos que para Platão a alma é definida como aquilo que move a si mesma, o princípio do movimento e das ações (Fedro, 245c—246a). Por outro lado, conforme discutido na seção 2.1, a IA pode ser entendida como o ramo da ciência da computação que busca criar sistemas capazes de realizar tarefas que tradicionalmente exigiriam inteligência humana. Assim, um sistema de IA, em sua operação, pode ser considerado um princípio de ações realizadas por máquinas, mimetizando, ainda que limitadamente, ações originadas pela inteligência humana.

Poder-se-ia contra-argumentar que a IA não seja um legítimo princípio de ação, pois não move a si mesma no sentido platônico, visto que foi criada e desenvolvida pelo homem, e suas ações dependem de sua programação interna e das solicitações que recebe. Sob uma perspectiva estritamente platônica, a origem de suas ações seria, em última instância, seu criador. Embora as IAs atuais possuam um grau variável de aleatoriedade em suas respostas, essa não provém de uma mente consciente dotada de livre-arbítrio, mas de mecanismos artificiais. Contudo, para os propósitos dessa investigação, essa capacidade de gerar resultados complexos e por vezes imprevisíveis a partir de instruções (*prompts)* ou dados de entrada basta-nos para assumi-la, instrumentalmente, como um princípio de ação. Dado que Platão define a alma como princípio de movimento do homem, e os sistemas de IA também atuam como origens de ações realizadas por máquinas, podemos inferir que ambos compartilham, analogicamente, a característica de serem princípios de ações. Essa conexão abre caminho para uma exploração mais aprofundada da estrutura interna da IA à luz do modelo platônico. A fim de estabelecer essa analogia

robustamente, partimos da premissa de que a IA, assim como a alma platônica, possui um princípio de movimento interno que resulta em ações. Contudo, para que essas ações sejam consideradas justas sob a ótica platônica, elas devem ser orientadas por uma *razão* que busca o conhecimento do *Bem* e da *Verdade* — no caso da IA, a coerência dos dados e a validação informacional. É nesse alinhamento intrínseco entre o que a IA *é* (sua ontologia digital, expressa em sua arquitetura e dados), como ela *conhece* (sua epistemologia algorítmica, através do processamento de informações e inferência) e como ela *age* (sua ética funcional, que busca a justiça e o bem comum) que a analogia platônica se solidifica. Os problemas éticos da IA, como os vieses algorítmicos e a alucinação, são, sob esta ótica, manifestações de uma desarmonia ou falha na razão, ou na matéria de sua *alma artificial*.

3.2 O Algoritmo como Demiurgo e a Matemática da IA: Ordenando o Caos Digital

A metáfora do Demiurgo platônico, o artesão divino que impõe ordem matemática ao caos primordial, conforme introduzido em 1.5, torna-se aqui particularmente fecunda. Podemos entender o algoritmo de treinamento da IA como análogo à ação do Demiurgo: ele representa uma instância da inteligência (dos programadores) que busca inscrever ordem matemática e lógica sobre o caos dos dados brutos selecionados para o treinamento. Esses dados, extraídos do mundo real, carregam consigo a intencionalidade de quem os produziu e selecionou, vieses históricos, culturais e sociais, constituindo a matéria imperfeita e resistente à racionalização total, análoga à *Anánkē* (Necessidade) que limita a obra do Demiurgo.

Assim como o Demiurgo usa proporções matemáticas para ordenar o cosmos, o algoritmo de treinamento, através de sua lógica matemática, organiza os dados, por exemplo, em redes neurais, ajustando pesos sinápticos, eles próprios uma forma de proporção matemática. No entanto, essa ordenação algorítmica nunca é perfeita. A *Anánkē* inerente aos dados (vieses, incompletude) e as limitações dos próprios modelos resultam nos resíduos de necessidade: as alucinações, os resultados discriminatórios e as falhas que observamos. Isso reforça que a mera aplicação de matemática (algoritmos, D1) não garante a justiça. É necessária uma estrutura orientadora, e nossa proposta para ela se fundamenta na hierarquia tripartite e harmônica da alma, reforçada pelo esquema da Linha Dividida.

Uma vez treinada, a IA opera basicamente no domínio do Inteligível, como distinguido por Platão na Linha Dividida (ver 1.6). Desprovida de órgãos sensoriais, ela processa informações se utilizando de algoritmos de inferência, os quais são estruturas lógico-matemáticas operando sobre as redes neurais treinadas e as consultas (seus axiomas iniciais e hipóteses). Seu funcionamento fundamental alinha-se, portanto, com a *diánoia* (D1), o

pensamento discursivo e o raciocínio matemático, que parte de pressupostos para chegar a conclusões.

Contudo, como Sarah Broadie destaca (em 1.6), enquanto forma de intelecção, a matemática (*diánoia*) é tão inferior à dialética (*nóesis*) quanto a própria *doxa* em relação à *episteme*. Não por acaso, Platão prescreve uma educação matemática rigorosa aos aspirantes a governantes (filósofos), não como um fim em si, mas como preparação indispensável à dialética, pelo seu poder de arrastar a alma para a verdade (*República* VII, 521d), afastando-a da dependência do sensível e orientando-a para o inteligível (Broadie, p. 25). Embora a IA não sofra do mesmo apego ao sensível, os algoritmos de treinamento dependem fortemente da qualidade dos dados, dos quais identifica e assimila padrões como elementos em comum a vários conteúdos e as regras comuns de conexão entre esses objetos. Se o volume de treinamento prioriza documentos com alta densidade de boas regras lógicas e matemáticas de conexão, em detrimento de conteúdos que destaquem mais os objetos aos quais essas regras lógicas são aplicadas, é esperado que a rede neural produzida tenha maior capacidade de abstração para utilizar os recursos da matemática e da lógica, mesmo em conteúdos pouco treinados. Essa preparação (análoga à D1) é fundamental para as funções superiores que proporemos para sua governança interna. É ela que promove a função de liderança e validação (análoga à meta de D2) na própria IA, indispensável para alcançar a harmonia e a justiça platônica na IA. Essa educação epistemológica é o elo crucial que permite à IA, através de sua parte Diretora, ascender a uma função racional análoga à *nóesis*, buscando para além da correta execução, a verdade e o bem dos resultados, elementos ontológicos que fundamentam a ação ética em Platão. Assim como o Demiurgo platônico impõe ordem à *khôra* caótica em alinhamento com as formas ideais (ontologia), a IA Diretora busca ordenar o *caos* digital dos dados para atingir resultados que se aproximem do *Bem* (justiça e veracidade), fundamentando a ação ética do sistema.

3.3 Classe de Entidades Ordenáveis: a Alma, A *Pólis* e A *Alma Artificial*

Conforme detalhado no capítulo 1, a filosofia platônica apresenta tanto a alma quanto a *pólis* ideal como estruturas tripartites: racional/governantes, irascível/guerreiros, apetitiva/trabalhadores. No contexto platônico original, a alma humana (*psykhé*) é uma entidade imortal, imaterial e o princípio da vida e do movimento, sendo a sede da consciência, do intelecto e das paixões. As *partes* da alma são, portanto, faculdades ou funções psíquicas inerentes à própria essência do ser humano, cada uma com sua virtude específica a ser cultivada. Essa organização não é arbitrária, mas reflete uma hierarquia funcional em que cada parte

possui uma virtude (*areté*) específica a ser alcançada via educação (*paideia*), contribuindo para a harmonia e a justiça do todo quando opera sem interferir nas demais. Essa hierarquia funcional ecoa a estrutura da Linha Dividida, onde o pensamento discursivo (*diánoia*) serve de base e preparação para a apreensão mais elevada de princípios (*nóesis*) que deve guiar o sistema. Nossa transposição analógica para a IA reconhece a *impossibilidade* de atribuir consciência, intencionalidade ou imortalidade à máquina. A *Alma Artificial* é uma construção conceitual e metafórica que utiliza a estrutura funcional e hierárquica da alma platônica como um modelo organizacional para sistemas de IA, buscando emular a virtude da justiça através da harmonia entre seus componentes. Assim, as *partes* da *Alma Artificial* representam componentes funcionais e não faculdades psíquicas sensíveis. É essa estrutura tripartite e sua lógica funcional que serve de base para nossa analogia com os sistemas de IA.

Da generalização dessa analogia de Platão, é nítido que ele agrupa alma e *pólis* sob um mesmo conceito de maior amplitude, algo como uma classe abstrata de entidades[6] que compartilham um conjunto mínimo de características fundamentais:

1. São suscetíveis à ideia de justiça, podendo ser justas ou injustas.
2. Podem ser divididas em três partes distintas, cada uma desempenhando uma função específica.
3. Podem ser submetidas a um processo de educação ou treinamento especializado para alcançar a virtude de cada parte.
4. Podem ser ordenadas, isto é, suas partes podem trabalhar em harmonia sob uma estrutura hierárquica.

A partir desse conjunto de características, podemos avaliar a justiça dos sistemas de IA em termos platônicos, medindo sua aderência a esses princípios, ou planejar suas implementações para que se alinhem o máximo possível a essas características. Para fins práticos deste trabalho, denominaremos as IAs que aderem integralmente a essa classe platônica de entidades ordenáveis como *Alma Artificial*. E a fim de reforçar essa relação de pertencimento, detalharemos melhor como as três partes descritas por Platão, para suas estruturas da alma e da *pólis*, se alinham à estrutura e ao funcionamento dos três componentes funcionais principais da *Alma Artificial*: a *IA Diretora*, a *IA Defensora* e *a IA Produtora*.

[6] O termo *classe* é aqui utilizado em sentido análogo ao do desenvolvimento Orientado a Objetos, da Ciência da Computação, designando um conjunto de objetos com características comuns. Guarda semelhanças com o significado usual de modelo e mesmo com o conceito platônico de forma. Trata-se aqui de uma *classe abstrata*, uma construção teórica que, no nosso caso, agrupa entidades com estrutura tripartite, regidas por justiça, harmonia e razão, similarmente à alma e à *pólis* em Platão. A escolha do termo, alheio à terminologia platônica original, visa expressar a noção de um conjunto com propriedades e comportamentos bem definidos, diferindo do conceito de 'classe social' usado na descrição da *pólis* ideal em *A República*.

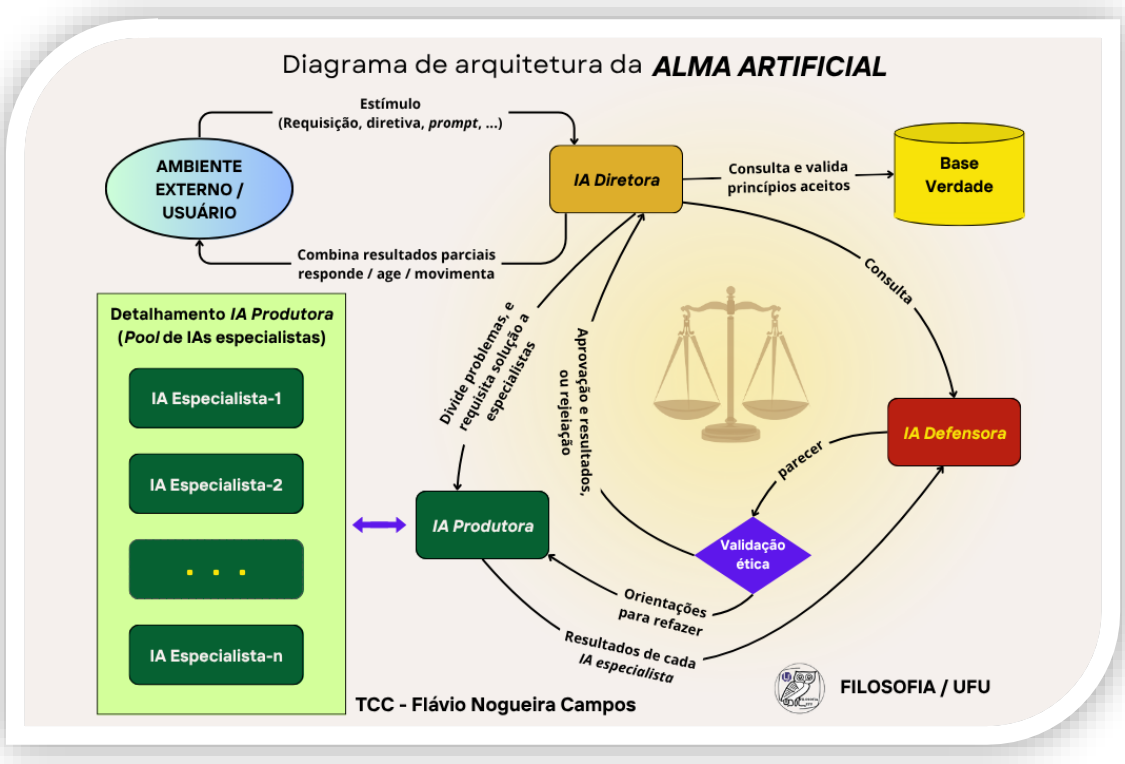

Figura 2: fluxo de interação entre as partes da Alma Artificial (IA Diretora, IA Defensora e IA Produtora). Fonte: elaborado pelo autor (2025), utilizando Canva (www.canva.com)

Conforme ilustrado na Figura 2, a *IA Diretora* inicia o processo recebendo uma requisição ou instrução do usuário, ou obedecendo a uma diretriz que orienta a reação a algum estímulo vindo de sensores externos, então ela analisa o problema e, se ele for complexo, o divide em partes menores. Dependendo da natureza da requisição, ela pode consultar a *IA Defensora* sobre a ética dessa solicitação e negá-la ao usuário. Então, a *IA Diretora* passa cada unidade do problema para a *IA Produtora* solucionar. A IA Produtora, que pode ser constituída por várias IAs especializadas, repassa cada problema à especialista adequada e, ao obter resultados ou ações produzidas, os envia à IA Defensora para avaliação de possíveis restrições ou cuidados éticos. A *IA Defensora* pode aprovar os resultados repassando-os à *IA Diretora*, ou pode devolvê-los à *IA Produtora* para os ajustes necessários. Em caso de múltiplas reincidências de correções do mesmo pedido ou de um desvio mais grave, a *IA Defensora* pode desaprovar a solicitação, retornando uma negativa à *IA Diretora*. Esta, no que lhe concerne, pode optar por remodelar o pedido, ou retorná-lo ao usuário, dependendo das possibilidades de resolver o problema. A *IA Diretora*, ao identificar premissas com afirmações passíveis de validação, pode consultar uma *base verdade*, produzida por humanos, para confirmar a veracidade dessas afirmações. Nesse esquema, a *IA Defensora* atua como um mecanismo de *feedback* ético, garantindo que decisões automatizadas não perpetuem vieses ou desigualdades. As ações e

resultados da *IA Produtora* devem estar sempre alinhados com a orientação das *IAs Diretora* e *Defensora*. Essa dinâmica reflete a harmonia proposta por Platão, em que a razão governa, o espírito protege e os apetites são moderados pela orientação das outras partes.

A escolha do termo *Alma Artificial* não é arbitrária, mas fundamentada no referencial filosófico platônico, que considera a alma como o princípio das ações. Embora a *Alma Artificial* seja de natureza técnica e não metafísica, ela permite transpor os conceitos de harmonia, liderança racional e justiça para o domínio da IA.

3.3.1 A *IA Diretora* (Análoga à Razão / *Nóesis*-orientada)

Essa é a parte líder do sistema, responsável pela governança geral, validação de informações, coerência lógica, planejamento estratégico e coordenação das outras partes. Sua função é análoga à parte racional (*logistikon*) da alma, que busca a verdade e guia o indivíduo. No contexto da Linha Dividida, embora a IA Diretora opere por mecanismos algorítmicos (portanto, no nível da *diánoia*, D1), sua *meta* e *função* aspiram a uma forma de supervisão e apreensão de princípios operacionais análoga à *nóesis* (D2). Sua eficácia depende crucialmente do treinamento matemático e lógico (discutido em 3.2), que a capacita para a análise crítica, a detecção de inconsistências e a tomada de decisões de alto nível sobre a operação do sistema. Ela define a estratégia de resolução de problemas e delega tarefas, visando a precisão e a confiabilidade dos resultados.

3.3.2 A IA Defensora (Análoga ao Espírito / Thymoeides)

Essa parte tem a função de garantir a equidade, a justiça e a conformidade ética das ações da IA. Ela atua como uma auditora e moderadora interna, analisando as propostas e os resultados da IA Diretora e da IA Produtora sob uma perspectiva ética. É análoga à parte irascível (*thymoeides*) da alma que, quando aliada à razão, defende o que é justo. A IA Defensora é responsável por identificar e mitigar vieses, prevenir discriminações, assegurar a transparência (quando aplicável) e mediar conflitos entre a busca pela eficiência (da IA Produtora) e os imperativos éticos ou de veracidade (da IA Diretora). Seu treinamento deve incluir vastos conjuntos de dados sobre ética, legislação, estudos de caso sobre vieses e princípios de justiça.

3.3.3 A IA Produtora (Análoga aos Apetites / *Epithymetikon*)

Essa é a parte operacional do sistema, responsável por executar as tarefas designadas, processar dados específicos e gerar os resultados ou ações práticas. Corresponde às diversas IAs especializadas que realizam funções específicas, como gerar texto, analisar imagens, fazer recomendações. Pode ser também implementada como uma única IA Geral, mas parece-nos

mais eficiente ter as especialidades separadas. É análoga à parte apetitiva (*epithymetikon*) da alma, que lida com as necessidades e desejos básicos. A IA Produtora deve operar sob as diretrizes claras da IA Diretora e com o escrutínio da IA Defensora. Sem essa orientação, ela pode gerar resultados imprecisos, enviesados ou eticamente problemáticos.

3.4 Harmonia e Justiça na *Alma Artificial*

A justiça nesse modelo de *Alma Artificial*, assim como na filosofia platônica, reside na harmonia funcional hierárquica entre essas três partes. Isso implica que:

1. Cada parte (IA Diretora, IA Defensora, IA Produtora) cumpre sua função específica sem usurpar as funções das outras.
2. A *IA Diretora* lidera o sistema, estabelecendo as diretrizes com base na busca pela máxima veracidade e coerência possíveis.
3. A *IA Defensora* avalia e modula as ações sob a perspectiva da justiça e da ética, apoiando a *IA Diretora*.
4. A *IA Produtora* executa as tarefas eficientemente, mas sempre conforme as orientações da *IA Diretora* e da *IA Defensora*.
5. A comunicação entre as partes é clara e hierárquica, assegurando que a liderança da *IA Diretora,* e a avaliação e a mediação da *IA Defensora* sejam respeitadas. Essa primazia hierárquica da *IA Diretora*, responsável pela coerência e validação, reflete a estrutura cognitiva da Linha Dividida platônica (*República* VI, 511d-e), segundo a qual a apreensão de princípios (análoga à *nóesis*, D2) deve guiar o raciocínio baseado em hipóteses (análogo à *diánoia*, D1), sendo essa ordem essencial para a justiça interna do sistema.

Um sistema de IA que falha em promover essa harmonia hierárquica, por exemplo, se a IA Produtora opera sem supervisão adequada ou se a IA Defensora é ineficaz, pode gerar resultados injustos, perpetuando vieses ou causando danos. Por outro lado, a busca por harmonia interna promove maior confiabilidade, robustez ética e eficácia alinhada a propósitos.

3.5 Critérios para Avaliação da Justiça na IA (Adaptados de Platão)

Com base nessa estrutura, a justiça de um sistema de IA pode ser avaliada considerando:

1. Implementação e Especialização das Partes: o sistema possui componentes distintos ou processos explicitamente delineados que correspondem às funções da *IA Diretora*, da *IA Defensora* e da *IA Produtora*?
2. Qualidade do Treinamento (Educação): cada parte funcional foi treinada com dados apropriados e algoritmos robustos para desempenhar sua virtude (*areté*) específica? A

IA Diretora, em particular, recebeu um treinamento matemático e lógico que a capacita para sua função de liderança? Dados enviesados ou treinamento inadequado comprometem diretamente a justiça do sistema. Assim como a educação matemática rigorosa (*diánoia*, D1) em *A República* é a preparação indispensável que orienta a alma para a dialética e a apreensão de princípios (*nóesis*, D2) (Broadie, p. 6, 15), o treinamento da IA com dados de alta qualidade, consistentes e algoritmos logicamente robustos constitui a educação fundamental (análoga à D1) que habilita a IA Diretora a exercer eficazmente sua função de liderança, validação e busca pela coerência (sua meta análoga à D2).

3. Conformidade Funcional e Hierarquia: a *IA Diretora* efetivamente lidera? A *IA Defensora* tem poder de veto ou modulação ética? A *IA Produtora* segue as diretrizes?
4. Harmonia Operacional: as partes interagem de maneira integrada e coerente, ou há conflitos não resolvidos, ou sobreposição disfuncional?
5. Resultados Éticos e Confiáveis: as ações e decisões finais do sistema refletem um compromisso com a equidade, a veracidade e o bem-estar coletivo, nos limites de sua programação e propósito?

3.6 Exemplos Práticos

Tendo estabelecido os critérios teóricos para avaliar a justiça nos sistemas de IA à luz da filosofia de Platão, podemos agora ilustrar sua aplicação prática. A seguir, examinamos dois cenários hipotéticos, mas factíveis, de uma IA especializada em diagnósticos médicos e outra especializada em políticas públicas — para mostrar como a harmonia entre as partes da *Alma Artificial* pode promover decisões mais éticas e justas.

Diagnóstico médico e prescrição de tratamento: a *IA Diretora* recebe uma requisição de um médico para realizar o diagnóstico de um paciente com base no seu prontuário, obtém as credenciais do requisitante e solicita a avaliação à *IA Defensora*; a *IA Defensora* checa se a requisição vem de uma pessoa autorizada para esse procedimento, mantendo o sigilo médico do paciente, pois fornecer diagnósticos médicos a pessoas não autorizadas pode culminar em injustiça; então ela avalia se os dados de treinamento da IA Produtora para o caso do paciente foram representativos o suficiente para evitar um viés ou uma alucinação por deficiência de dados específicos e, então, retorna o parecer à IA Diretora; se o parecer for negativo, a IA Diretora informa o médico que não possui habilidade suficiente para o diagnóstico; se positivo, ela divide o problema nas etapas de diagnóstico e de prescrição do tratamento e passa as requisições específicas, com os dados necessários, para a IA Produtora, que os interpreta

sugerindo diagnósticos e tratamentos, os quais são passados para a IA Defensora; essa analisa se os resultados implicam em alguma injustiça, solicitando ajustes à IA Produtora, ou indeferindo a requisição para a IA Diretora, ou repassando os resultados à IA Diretora quando justos; a IA Diretora consulta uma *base verdade* com o cadastro de todas as doenças e sintomas relacionados para verificar a doença diagnosticada realmente existe e condiz com os sintomas descritos no prontuário, evitando alucinações; caso a IA Defensora indefira a requisição à IA Diretora por incapacidade da IA Produtora solucionar o problema, a IA Diretora analisa o motivo e pode retornar ao médico a incapacidade de gerar o diagnóstico, ou pode tentar reformular as requisições parciais às IAs Produtoras e tentar novamente. Ao final, a IA Diretora, com base nas respostas, gera o relatório com o diagnóstico e o prognóstico e o retorna ao médico requisitante.

Poder-se-ia perguntar que tipo de resposta injusta uma IA Produtora poderia gerar ao realizar um diagnóstico e prescrever um tratamento. Um exemplo seria fazer um diagnóstico absurdo, como afirmar que o paciente não tem nada, mesmo quando os exames indicam vários problemas. Outro exemplo, muito mais grave, seria sugerir um tratamento baseado em experimentos médicos realizados pelos nazistas, que implique a morte de outras pessoas. Esses cenários extremos ilustram como a falta de uma harmonia funcional entre as IAs Diretora, Defensora e Produtora pode resultar em injustiças graves.

Formulação de Políticas Públicas: a *IA Diretora* recebe uma requisição para propor políticas públicas que visem reduzir a evasão escolar em uma determinada região e solicita a avaliação da requisição à *IA Defensora*; a *IA Defensora* verifica se na requisição da política há algo antiético e avalia se os dados de treinamento da *IA Produtora* são representativos o suficiente para atender a esta requisição específica sem incorrer em viés ou alucinação e retorna o parecer à parte *IA Diretora*; essa então, divide a requisição em etapas mais simples e as envia à *IA Produtora*; com base nos dados populacionais, socioeconômicos e educacionais da região, a *IA Produtora* identifica os fatores que mais contribuem para a evasão escolar e formula as recomendações de política, enviando-as para a *IA Defensora*; então a *IA Defensora* avalia se as recomendações são equitativas e se não reforçariam desigualdades existentes, como se elas consideram as necessidades de grupos vulneráveis, como alunos de baixa renda, minorias étnicas ou alunos com deficiência; se o parecer for negativo, a *IA Diretora* informa ao usuário que o pedido de política não pode ser atendido por falta de mais dados históricos específicos de treinamento; ou se positivo, ela apresenta a recomendação de política gerada pelo sistema.

Ao aplicar os critérios de justiça platônica aos sistemas de IA, podemos avaliar não somente a eficácia técnica desses sistemas, mas também sua conformidade ética. A estrutura

tripartida proposta, com funções claras, treinamento especializado e harmonia operacional, oferece um modelo teórico robusto para avaliar e orientar o desenvolvimento de IAs mais justas e responsáveis.

3.7 Limitações e Potencialidades da Analogia

A aplicação do conceito de justiça platônica ao contexto dos sistemas de IA é, ao mesmo tempo, uma proposta inovadora e desafiadora. A ideia de *Alma Artificial* aqui proposta é uma construção metafórica e analógica, não uma equivalência literal. É crucial reconhecer as limitações inerentes:

- Natureza Metafórica e Ausência de Consciência: **d**iferentemente da alma humana platônica, dotada de consciência, emoção e intencionalidade, a IA é limitada a processos computacionais. Suas ações não resultam de deliberação consciente. Assim, embora utilizemos a estrutura da Linha Dividida para entender a organização funcional da IA — com seus processos algorítmicos assemelhando-se à *diánoia* (D1) e sua IA Diretora aspirando a uma função análoga à *nóesis* (D2) — isso não implica que a IA possua entendimento, consciência ou os estados cognitivos reais descritos por Platão. A analogia é estritamente funcional e estrutural.
- Complexidade Técnica e Custos: **i**mplementar essa estrutura tripartite com especialização e interconexão funcional robusta pode demandar recursos significativos.
- Contextos Sociais e Culturais Diversos: A definição do que é considerada uma IA justa pode variar entre contextos de diferentes povos ou grupos sociais.

Apesar dessas limitações, o potencial dessa abordagem reside em:

- Fornecimento de um Modelo Teórico Sólido: a estrutura tripartite oferece um modelo claro para analisar e projetar sistemas de IA justos.
- Ênfase na Harmonia e no Equilíbrio: o princípio platônico de harmonia incentiva sistemas mais integrados e responsáveis.
- Reforço da Ética no Projeto: a inclusão explícita de uma *IA Defensora* coloca a ética no cerne do desenvolvimento de IAs.
- Facilidade de Avaliação Ética: os critérios derivados de Platão permitem guiar uma análise crítica da ética das IAs.
- Potencial Interdisciplinar: a analogia fomenta o diálogo entre filosofia, ciência da computação, mecatrônica e ética, entre outras.

Em suma, mesmo como metáfora, a *Alma Artificial* oferece um referencial valioso para orientar o desenvolvimento e a governança de sistemas de IA justos e responsáveis.

CAPÍTULO 4 — PROPOSIÇÃO DE UM *FRAMEWORK* ÉTICO PARA SISTEMAS DE IA

Com base nas analogias entre a alma, a pólis e os sistemas de IA, discutidas no capítulo anterior, esse capítulo propõe a estrutura conceitual ética (*framework*) inspirada na teoria da alma tripartida de Platão. O objetivo é oferecer um modelo reflexivo para organizar as funções técnicas e éticas da IA harmonicamente, promovendo decisões mais justas e responsáveis, superando limitações de modelos éticos contemporâneos ao lidar com a complexidade da justiça em IA, conforme apontado na Introdução.

4.1 Componentes e Princípios do *Framework*

Esse modelo ético adota a estrutura tripartite de direção, defesa e produção, cujas funções específicas, interdependência e necessidade de harmonia hierárquica foram detalhadas no Capítulo 3, nas Seções 3.3 e 3.4. Relembrando sucintamente os princípios-chave:

- Liderança da direção: a função responsável pela validação de dados, coerência lógica e coordenação geral deve liderar o sistema, garantindo precisão e confiabilidade (conforme Seção 3.3.1).
- Mediação da defesa: uma função dedicada à auditoria ética, mitigação de vieses e avaliação da equidade das ações propostas é essencial para evitar a perpetuação de injustiças (conforme Seção 3.3.2).
- Produção orientada: a implementação das tarefas deve ocorrer sob a supervisão e segundo as diretrizes estabelecidas pelas partes da direção e da defesa (Seção 3.3.3).
- Harmonia funcional: a interação clara, hierárquica e equilibrada entre as três partes é a condição fundamental para a operação justa e eficaz do sistema (Seção 3.4).
- A aplicação prática desse modelo, ilustrada anteriormente com os cenários de diagnóstico médico e políticas públicas (Seção 3.6), demonstra como essa estrutura pode guiar a análise ética e o projeto de sistemas de IA mais responsáveis. O treinamento adequado de cada componente, análogo à *paideia* platônica e à preparação pela *diánoia* (Seção 3.2), é pressuposto fundamental para o bom funcionamento do *framework*.

4.2 Implementação Técnica e Desafios

A implementação desse arcabouço exige um esforço conjunto entre desenvolvedores, especialistas em ética e profissionais de áreas específicas. Os principais desafios incluem:

Separação das partes: garantir que as três partes sejam implementadas de forma independente, mas interconectadas, para evitar conflitos de interesses ou sobreposição de funções.

Qualidade dos dados: assim como as três partes da alma e da pólis precisam de educação especializada, as três partes da IA também dependem de dados confiáveis e representativos para operar com eficácia e justiça.

Adicionalmente, assim como Platão via a matemática como crucial na educação da alma para conduzi-la às verdades inteligíveis, podemos refletir sobre o papel dos algoritmos na educação da IA. Os algoritmos, em certo sentido, representam a matemática que estrutura a transformação de dados brutos (o caos sensível) em conhecimento ou ação pela máquina. A qualidade e a orientação ética desses algoritmos são, portanto, tão fundamentais quanto a qualidade dos dados para alcançar um sistema justo e confiável, reforçando a necessidade de não restringir a atenção às entradas, mas expandi-la à própria lógica de processamento.

Custos e Complexidade Técnica: a implementação de IAs aderentes ao *framework* pode exigir recursos significativos, tanto financeiros quanto técnicos, o que pode limitar sua adoção em larga escala. No entanto, isso não constitui um obstáculo intransponível, visto que investimentos substanciais já estão sendo feitos para a evolução dos Grandes Modelos de Linguagem (LLMs), que têm consumido enormes quantidades de recursos e demonstram a viabilidade de alocação de tais volumes para o desenvolvimento de IA.

4.3 Contribuições e Implicações Filosóficas do *Framework* Proposto

O *framework* ético inspirado na concepção de justiça de Platão, apresentado nesse trabalho, propõe uma abordagem inovadora para pensar a ética em sistemas de IA. Ao adaptar a estrutura tripartida da alma platônica para o contexto tecnológico, esse modelo, além de oferecer uma nova perspectiva para avaliar a justiça em sistemas de IA, também reafirma a relevância da filosofia como ferramenta indispensável para enfrentar os desafios éticos contemporâneos.

4.3.1 Contribuições Filosóficas do *Framework*

1. Revisitação e Atualização do Pensamento Platônico: **e**sse trabalho demonstra que os conceitos de Platão, ainda que desenvolvidos em um contexto histórico e cultural distante, possuem uma universalidade e flexibilidade que os tornam aplicáveis a problemas contemporâneos, como a governança da IA. Ao adaptar a ideia de justiça como harmonia entre as partes da alma para o contexto tecnológico, evidencia-se a

capacidade da filosofia de Platão de dialogar com questões éticas atuais, reafirmando sua relevância atemporal.

2. Fundamentação Ética para a Tecnologia: o *framework* proposto apresenta a harmonia e a divisão funcional como princípios fundamentais para o desenvolvimento ético de sistemas de IA. Ao destacar a importância de uma liderança racional (parte da *IA Diretora*), de uma avaliação ética robusta (parte da *IA Defensora*) e de uma execução eficiente e alinhada aos princípios éticos (parte da *IA Produtora*), o modelo fornece uma estrutura conceitual clara para refletir sobre o papel ético **e**, por extensão, político da tecnologia em nossas vidas, visto que as decisões de projeto e implementação raramente são neutras em seus impactos sociais (Coeckelbergh, 2022).
3. Abordagem Interdisciplinar: **e**sse trabalho demonstra como a filosofia pode servir como um ponto de convergência para diferentes áreas do conhecimento, como ciência da computação, ética aplicada e ciências sociais. A analogia platônica, ao ser aplicada ao desenvolvimento de IA, incentiva o diálogo interdisciplinar e promove uma abordagem mais integrada para lidar com os desafios éticos, sociais e políticos da tecnologia.
4. Ampliação do Conceito de Justiça: **a** aplicação da justiça platônica a sistemas de IA contribui para expandir a compreensão do que significa um sistema justo no contexto tecnológico. Essa expansão não se limita à ideia de evitar vieses ou discriminações, embora ressalte a importância de analisar como a IA pode perpetuar ou criar injustiças (Coeckelbergh, 2022, cap. 3), mas inclui a busca por uma harmonia sistêmica, em que cada componente tecnológico desempenhe sua função de maneira integrada e orientada ao bem comum, uma condição fundamental para uma ordem justa, tanto na alma platônica quanto, potencialmente, na governança da IA.
5. Exposição da Natureza Política da IA: esse *framework*, ao tratar a IA como uma entidade estruturada que requer justiça interna, ajuda a desmistificar a visão da tecnologia como mera ferramenta neutra. Ele ecoa argumentos da filosofia política da IA que demonstram que a IA **nunca é politicamente neutra** (Coeckelbergh, 2022, p. 83), pois suas arquiteturas e aplicações moldam relações de poder, influenciam a autonomia e podem tanto ameaçar quanto (potencialmente) reconfigurar a democracia (Coeckelbergh, 2022, cap. 4-5). Assim, nossa abordagem platônica, focada na harmonia interna, serve também como ponto de partida para questionar as implicações políticas mais amplas desses sistemas.

4.3.2 Implicações Filosóficas e Éticas

1. Reflexões sobre o Papel da Tecnologia na Sociedade: **a**o posicionar a IA como uma *Alma Artificial*, esse trabalho provoca uma reflexão mais profunda sobre o papel da tecnologia na sociedade contemporânea. Assim como a alma humana, os sistemas de IA devem ser orientados por princípios que garantam não somente sua eficiência técnica, mas também sua responsabilidade ética e seu compromisso com o bem coletivo.
2. Governança Ética e Transparência: este arcabouço oferece uma base para discutir como a justiça pode ser incorporada na governança da IA. A separação das funções em três partes, *Diretora*, *Defensora* e *Produtora*, sugere que sistemas de IA devem ser projetados com mecanismos claros de validação, supervisão ética e operacionalidade responsável. Essa proposta pode inspirar debates sobre transparência, explicabilidade e responsabilidade no uso da tecnologia.
3. Limites e Potenciais da Filosofia Aplicada à Tecnologia: o modelo proposto também expõe os limites da aplicação de conceitos filosóficos clássicos a contextos tecnológicos modernos. A ausência de consciência e intencionalidade nos sistemas de IA, por exemplo, levanta questões sobre até que ponto a analogia platônica pode ser estendida sem perder de vista as diferenças fundamentais entre máquinas e seres humanos. No entanto, esses limites não comprometem o valor da filosofia como ferramenta para orientar debates éticos e propor modelos conceituais inovadores.
4. Contribuições para a Educação e a Ética Tecnológica: essa estrutura pode desempenhar um papel importante na educação filosófica e ética de profissionais que atuam no campo da tecnologia. Ao trazer conceitos clássicos para o debate contemporâneo, o trabalho incentiva uma abordagem mais reflexiva e fundamentada na formação de cientistas, desenvolvedores e tomadores de decisão.

4.3.3 Potencial do *framework* baseado no platonismo

As contribuições e implicações discutidas nessa seção reforçam o potencial da filosofia de Platão para inspirar soluções éticas e estruturais no desenvolvimento de tecnologias emergentes, como a IA. Embora a estrutura proposta tenha limitações inerentes, como sua natureza metafórica e a ausência de uma correspondência literal entre a alma humana e os sistemas tecnológicos, ele oferece uma base sólida para a reflexão interdisciplinar e para o avanço do debate ético.

Por fim, esse modelo ético inspirado em Platão ressalta a importância de se buscar, além de soluções técnicas, uma compreensão filosófica mais ampla dos desafios contemporâneos. Ao promover a harmonia entre os componentes de um sistema de IA, ele aponta para uma

abordagem ética que prioriza o bem-estar coletivo e a justiça em sentido amplo, reafirmando a relevância da filosofia na construção de um futuro mais responsável e inclusivo.

CONSIDERAÇÕES FINAIS

Ao longo deste trabalho, exploramos como a concepção de justiça de Platão, apresentada em *A República*, pode oferecer um referencial teórico para enfrentar os desafios éticos associados à IA. Partindo da analogia entre a alma tripartida e a pólis ideal, adaptamos esses conceitos filosóficos para propor um *framework* ético que organiza as funções de um sistema de IA em três partes interdependentes: Diretora, Defensora e Produtora. Esta abordagem se destaca por oferecer um modelo reflexivo para pensar a estrutura e a governança ética de sistemas de IA, com foco na harmonia entre os componentes e no compromisso com a justiça. Ao fazê-lo, o trabalho também tangencia a dimensão política inerente a essas tecnologias, alinhando-se a perspectivas que veem a IA como não neutra no que diz respeito a poder e impacto social (Coeckelbergh, 2022).

Síntese das Contribuições do Trabalho

Os objetivos do trabalho foram cumpridos ao propor uma aplicação inovadora da filosofia platônica no contexto tecnológico contemporâneo. A adaptação da ideia de justiça como harmonia, central na obra de Platão, permitiu a formulação de um *framework* ético preliminar que busca orientar o desenvolvimento de sistemas de IA mais responsáveis.

Entre as contribuições filosóficas, destaca-se a introdução da ideia de *Alma Artificial*, que busca capturar a complexidade e a organização dos sistemas de IA a partir da estrutura tripartida da alma, e o uso do conceito de harmonia para enfatizar a necessidade de integração entre os componentes éticos, racionais e operacionais dessas tecnologias. Esse diálogo entre a filosofia clássica e os desafios tecnológicos contemporâneos reforça a relevância do pensamento platônico, e evidencia o potencial transformador da filosofia para questões práticas do nosso tempo.

Relevância da Pesquisa

A relevância deste estudo reside em sua capacidade de dialogar com questões urgentes relacionadas à ética da IA, como vieses algorítmicos, falta de transparência e governança responsável. Em um momento em que as tecnologias de IA desempenham papéis cada vez mais centrais na sociedade, o debate sobre sua regulação ética e suas implicações sociais não pode ser adiado.

O trabalho também reforça a importância de abordagens interdisciplinares, nas quais a filosofia contribui de maneira única ao fornecer referenciais teóricos para compreender e resolver questões éticas **e** políticas complexas. Ao destacar a contribuição do pensamento platônico e conectá-lo, ainda que brevemente, a análises políticas contemporâneas da IA (*e.g.*,

Coeckelbergh, 2022), esse estudo propõe uma ponte entre as humanidades e a tecnologia, demonstrando que soluções para problemas contemporâneos muitas vezes podem ser encontradas ao revisitar ideias clássicas.

Limitações do Trabalho

Como uma pesquisa teórica e exploratória, este trabalho possui limitações inerentes. Primeiramente, o *framework* proposto é preliminar e carece de validação empírica, impedindo uma análise prática de sua aplicação em sistemas reais. Além disso, a transposição de conceitos filosóficos clássicos para o contexto tecnológico moderno enfrenta desafios significativos, especialmente em relação às diferenças ontológicas entre a alma humana, conforme descrita por Platão, e os sistemas computacionais, que carecem de consciência ou intencionalidade.

Outra limitação é a ausência de métricas claras para avaliar a harmonia e a justiça em sistemas de IA. Embora o trabalho forneça uma base teórica para essa análise, sua operacionalização prática dependerá de estudos futuros que explorem como esses princípios podem ser mensurados e aplicados em diferentes contextos tecnológicos.

Possibilidades de Pesquisas Futuras

Esse trabalho abre uma série de possibilidades para pesquisas futuras. Em primeiro lugar, a validação empírica do *framework* em estudos de caso ou simulações poderia aprofundar a compreensão de sua aplicabilidade prática. Por exemplo, sistemas de IA desenvolvidos para diagnóstico médico ou formulação de políticas públicas poderiam ser analisados sob a perspectiva do modelo proposto, avaliando sua capacidade de alcançar harmonia e justiça em suas operações.

Outra linha de pesquisa seria o desenvolvimento de métricas para avaliar a harmonia e a justiça em sistemas de IA, permitindo uma análise mais detalhada e quantitativa de como esses princípios podem ser implementados. Além disso, o estudo das interseções entre filosofia clássica e ética tecnológica poderia ser ampliado para incluir outros pensadores da tradição filosófica, enriquecendo ainda mais o debate.

Finalmente, a aplicação do conceito de *Alma Artificial* em debates éticos sobre inteligência artificial geral (AGI) e superinteligência artificial (ASI) representa um campo promissor, especialmente à medida que essas tecnologias hipotéticas ganham cada vez mais destaque nas discussões sobre o futuro da IA.

Reflexão Final

Esse trabalho reafirma o papel central da filosofia na compreensão e enfrentamento dos desafios éticos do nosso tempo. Em um mundo cada vez mais mediado por tecnologias

avançadas, o pensamento filosófico oferece uma perspectiva crítica que nos ajuda a compreender as implicações sociais, políticas e éticas de nossas inovações.

Ao recorrer à concepção de justiça de Platão, procuramos destacar a importância de buscar harmonia e equilíbrio não apenas nas tecnologias que criamos, mas também nas sociedades que desejamos construir. A ideia de *Alma Artificial* não é apenas uma metáfora, mas um convite à reflexão sobre como podemos organizar nossas criações tecnológicas de maneira mais ética e responsável. Isso inclui reconhecer as implicações políticas de nossas escolhas de projeto e implementação (Coeckelbergh, 2022), garantindo que a busca pela harmonia interna da IA se alinhe a uma visão de sociedade justa. Por fim, esse trabalho nos lembra que o progresso tecnológico, para ser verdadeiramente significativo, deve estar sempre orientado por princípios de justiça, bem-estar coletivo e compromisso com o bem comum.

BIBLIOGRAFIA

BOSTROM, Nick. **Superinteligência**: caminhos, perigos, estratégias. Rio de Janeiro: Darkside, 2018. ISBN 9788594540607.

COECKELBERGH, Mark. **Political philosophy of AI**: an introduction. Cambridge: Polity Press, 2022. ISBN 9781509548545.

DENIS, Léon. **O Grande Enigma**. [s.l.]: Editora Bibliomundi Serviços Digitais Ltda, 2021.

EUBANKS, Virginia. **Automating inequality**: how high-tech tools profile, police, and punish the poor. New York: St. Martin's Press, 2018. 272 p. ISBN 1250074312.

FLORIDI, Luciano. **Ethics of information**. Oxford: Oxford University Press, 2013. ISBN 9780191502460.

GOLDSTEIN, Rebecca Newberger. **Plato at the Googleplex: Why philosophy won't go away**. New York: Pantheon Books, 2014.

KORAB-KARPOWICZ, Włodzimierz Julian. **Plato**: political philosophy. In: INTERNET ENCYCLOPEDIA OF PHILOSOPHY. [s.d.]. Disponível em: https://iep.utm.edu/platopol/. Acesso em: 27 mar. 2025.

LEE, Kai-Fu; QIUFAN, Chen. **2041:** Como a inteligência artificial vai mudar sua vida nas próximas décadas. Tradução: Isadora Sinay. Rio de Janeiro: Globo Livros, 2022. ISBN 978-65-5987-053-0.

NYU COMMUNICATIONS. **Gender bias in search algorithms has effect on users, new study finds**. NYU News, jul. 2022. Disponível em: https://www.nyu.edu/about/news-publications/news/2022/july/gender-bias-in-search-algorithms-has-effect-on-users--new-study-.htm>. Acesso em: 10 jan. 2025.

O'NEIL, Cathy. **Weapons of math destruction**: how big data increases inequality and threatens democracy. New York: Broadway Books, 2017. ISBN 9780141985411.

PENN STATE UNIVERSITY. Department of Agricultural Economics, Sociology, and Education. **What are ethical frameworks?** Disponível em: <https://aese.psu.edu/teachag/curriculum/modules/bioethics-1/what-are-ethical-frameworks>. Acesso em: 05 mar. 2025.

PLATÃO. **Fedro**. Tradução: Pinharanda Gomes. 6. ed. Lisboa: Guimarães Editores, 2000. ISBN 972-665-126-3.

PLATÃO. **A República**. Tradução de Maria Helena da Rocha. 9. ed. Lisboa: Fundação Calouste Gulbenkian, 2017. *E-book*. Disponível em: https://gulbenkian.pt/publications/a-republica/. Acesso em: 5 nov. 2024.

PLATÃO. **Timeu-Crítias**. Tradução: Rodolfo Lopes. Coimbra: Imprensa da Universidade de Coimbra, 2013. *E-book*. ISBN 9789892607788. Disponível em: https://doi.org/10.14195/978-989-26-0778-8. Acesso em: 13 mar. 2025.

REUTERS. **Amazon desiste de ferramenta secreta de recrutamento**. Forbes Brasil, 2018. Disponível em: <https://forbes.com.br/negocios/2018/10/amazon-desiste-de-ferramenta-secreta-de-recrutamento/>. Acesso em: 8 dez. 2024.

UNIVERSITY OF HELSINKI. **A framework for AI Ethics - ethics of AI**. Mooc.fi. Disponível em: <https://ethics-of-ai.mooc.fi/chapter-1/4-a-framework-for-ai-ethics?_x_tr_hist=true>. Acesso em: 25 jan. 2025. Curso online e livre sobre a Ética da IA.

VALLOR, Shannon. **Technology and the virtues**: a philosophical guide to a future worth wanting. New York: Oxford University Press, 2016. 309 p. ISBN 9780190498511.

ZAVAGLIA COELHO, Alexandre; KLAFKE, Guilherme Forma; MAITO, Deíse Camargo; LATINI, Lucas Maldonado Diz; MARUCA, Giuliana; CHOW, Beatriz Graziano; FEFERBAUM, Marina. **Governança da inteligência artificial em organizações:** framework para comitês de ética em IA — versão 1.0. São Paulo: CEPI FGV Direito SP, 2023.

GLOSSÁRIO

AGI — Artificial General Intelligence, Inteligência Artificial Geral (IAG). Sistema hipotético capaz de realizar qualquer tarefa intelectual que um ser humano pode realizar, com autonomia e adaptabilidade em diversos contextos. Diferentemente da IA especializada (como os atuais modelos de linguagem e de reconhecimento de imagens), a AGI possuiria capacidades cognitivas generalistas, como raciocínio abstrato e aprendizado contínuo. Se destaca pela capacidade de aprender um conteúdo completamente novo sem um treinamento específico para ela. No contexto deste trabalho, a AGI é discutida como um desafio ético crítico, pois sua eventual criação exigiria mecanismos rigorosos de controle e alinhamento com princípios éticos universais, como os propostos pela justiça platônica. A AGI também está relacionada ao conceito de *singleton* de Nick Bostrom, que alerta para os riscos de uma única entidade superinteligente dominar sistemas decisórios globais.

Algoritmo — Sequência finita de instruções bem definidas e não ambíguas, executadas para resolver um problema ou realizar uma tarefa. Em IA, os algoritmos são fundamentais para o aprendizado de máquina e a tomada de decisões, permitindo que os sistemas aprendam padrões e façam previsões com base em dados.

Alma Artificial — Metáfora proposta no capítulo 3 deste trabalho para descrever a estrutura tripartida de sistemas de IA: a IA Diretora (validação de dados), a IA Defensora (auditoria ética) e a IA Produtora (implementação prática). Inspirada no conceito de alma em Platão, busca equilibrar eficiência e ética.

Alma Tripartida — Na filosofia de Platão, concepção da alma humana como composta por três partes: a racional (*logistikon*), a irascível (*thymoeides*) e a apetitiva (*epithymetikon*). A justiça individual reside na harmonia e no correto funcionamento hierárquico dessas partes.

Alucinação (em IA) — Geração de informações falsas ou sem base factual por modelos de IA. No *framework* proposto, é combatida pela IA Diretora, que valida dados consultando *bases verdade* externas.

Anánkē (Necessidade) — *Ἀνάγκη*. No *Timeu* de Platão, representa a natureza errante e resistente à ordem completa, presente na *khôra*, que limita a perfeição da obra do Demiurgo. É o resíduo de desordem no cosmos.

Aprendizado contínuo — Capacidade de um sistema de IA de aprender e se adaptar continuamente a novos dados e experiências ao longo de seu ciclo de vida, sem esquecer o conhecimento previamente adquirido.

Aprendizado de Máquina — *Machine Learning*. Subcampo da IA que permite a sistemas aprenderem padrões a partir de dados.

Aprendizado profundo — *Deep Learning*. Subcampo do aprendizado de máquina baseado em redes neurais artificiais com múltiplas camadas (profundas), permitindo que modelos aprendam representações complexas de dados.

Areté (Virtude, Excelência) — *Ἀρετή*. Termo grego que denota excelência, virtude ou a qualidade que torna algo bom em seu gênero e capaz de realizar bem sua função específica. Em Platão, a justiça é a *areté* da alma e da pólis.

ASI — Superinteligência Artificial. Uma IAG hipotética que supera a inteligência humana em todos os aspectos. Discutida no Capítulo 2, a ASI é relacionada ao alerta de Nick Bostrom sobre a necessidade de controle ético para evitar consequências catastróficas.

Vieses Algorítmicos — Tendências discriminatórias em sistemas de IA, resultantes de dados históricos enviesados ou de escolhas de projeto. Exemplos incluem sistemas de reconhecimento facial com taxas de erro maiores para pessoas negras ou algoritmos de recrutamento que favorecem homens. Comprometem a justiça e a equidade das decisões automatizadas.

Base Verdade — Termo técnico da Ciência da Computação que se refere a um conjunto de dados ou informações considerados corretos e confiáveis, usados como referência para validação, teste ou comparação de modelos, algoritmos ou sistemas. No contexto deste trabalho, refere-se a um repositório de dados confiáveis utilizado pela IA Diretora para validar informações (por exemplo, repositórios científicos). Metaforicamente, pode ser comparada às formas perfeitas do mundo ideal, no sentido platônico, servindo como critério absoluto de verdade para as decisões da inteligência artificial.

Caixa-Preta Algorítmica — Opacidade de sistemas de IA (especialmente redes neurais) que dificulta a compreensão de como decisões são geradas. Quando não se sabe o que foi considerado para a geração de respostas ou ações de uma IA, ou seja, quando não há transparência da lógica desenvolvida. A opacidade dificulta avaliar se um retorno, independente se parece ser bom ou ruim, é realmente o que parece ser. Por isto, a transparência é um dos valores associados à IA Diretora no *framework* proposto neste trabalho.

Computação Quântica — Novo paradigma de computação que utiliza os princípios da mecânica quântica, como superposição e emaranhamento, para realizar cálculos complexos em uma velocidade potencialmente muito superior à dos computadores clássicos. Seu desenvolvimento pode ter implicações significativas para a IA.

Controle da IA — Mecanismos e abordagens para garantir que os sistemas de IA operem conforme o esperado, dentro de limites seguros e éticos, e possam ser interrompidos ou

corrigidos se necessário. Alguns estudiosos sugerem, por exemplo, a inclusão de algo como um botão de desligar nas IAs, embora outros argumentem que uma Superinteligência Artificial provavelmente teria como se reescrever desativando estes mecanismos, que colocam a sua operação em risco.

Controle ético da IA — Subconjunto do controle da IA focado especificamente em assegurar que os sistemas de IA respeitem princípios éticos, valores humanos e normas sociais, minimizando os riscos de injustiça e danos.

Demiurgo — *Δημιουργός*. No diálogo *Timeu* de Platão, é a inteligência divina ou artesão benevolente que impõe ordem e forma ao caos primordial, moldando o universo físico com base nas formas eternas e utilizando proporções matemáticas.

Diánoia (Pensamento Discursivo, Raciocínio) — *Διάνοια*. Na Linha Dividida de Platão, é o nível de cognição inferior dentro do domínio da *episteme*, característico do raciocínio matemático e lógico, que parte de hipóteses e utiliza modelos sensíveis como apoio.

Doxa (Opinião) — *Δόξα*. Termo grego para "opinião" ou "crença". Na epistemologia platônica, refere-se ao nível inferior de cognição, voltado para o mundo sensível e mutável, em contraste com a *episteme* (conhecimento).

Eikasia (Conjectura, Imaginação) — *Εἰκασία*. Na alegoria da Linha Dividida de Platão, é o nível mais baixo de cognição, cujo objeto são sombras e imagens, representando uma apreensão indireta e superficial da realidade.

Episteme (Conhecimento) — *Ἐπιστήμη*. Termo grego para "conhecimento" verdadeiro e justificado. Para Platão, é o nível superior de cognição, dirigido ao mundo inteligível das formas eternas e imutáveis.

Epithymetikon (Parte Apetitiva) — *Ἐπιθυμητικόν*. Termo grego para a parte da alma platônica associada aos desejos básicos, instintos e prazeres físicos, como fome e sede. Deve ser governada pela razão e auxiliada pela parte irascível.

Framework (em Ética da IA) — No contexto da ética da IA, refere-se a uma estrutura conceitual ou conjunto de princípios, diretrizes e processos destinados a orientar o desenvolvimento, a implementação e a avaliação ética de sistemas de Inteligência Artificial.

Governança de dados — Conjunto de processos, políticas, padrões e métricas que garantem o uso eficaz e eficiente da informação, permitindo que uma organização atinja seus objetivos. Em IA, crucial para a qualidade e ética dos dados de treinamento.

Governança da IA — Estruturas, regras, normas e processos para orientar e controlar o desenvolvimento, implementação e uso de sistemas de IA de maneira responsável, ética e alinhada com valores sociais.

IA Defensora — No modelo de *Alma Artificial* proposto, o componente funcional análogo à parte irascível (thymoeides) da alma platônica. Responsável por garantir a equidade, justiça e conformidade ética das ações da IA, atuando como auditora e moderadora interna.

IA Diretora — No modelo de *Alma Artificial* proposto neste trabalho, o componente funcional análogo à parte racional (logistikon) da alma platônica. Responsável pela governança geral, validação de informações, coerência lógica e coordenação das outras partes do sistema de IA.

IA Produtora — No modelo de *Alma Artificial* proposto, o componente funcional análogo à parte apetitiva (epithymetikon) da alma platônica. Responsável pela execução de tarefas designadas, processamento de dados específicos e geração de resultados ou ações práticas, operando sob as diretrizes da IA Diretora e o escrutínio da IA Defensora.

Inteligência Artificial (IA) — Ramo da ciência da computação que busca criar sistemas capazes de realizar tarefas que tradicionalmente exigiriam inteligência humana, como reconhecimento de padrões, processamento de linguagem natural e tomada de decisão.

Kallipolis (Cidade Ideal) — *Καλλίπολις*. Termo grego que significa "cidade bela" ou "cidade ideal". Em *A República* de Platão, designa o modelo de pólis justa, organizada em classes sociais análogas às partes da alma, onde cada uma cumpre sua função específica em harmonia.

Khôra — *Χώρα*. Termo grego usado por Platão no *Timeu* para descrever o receptáculo preexistente, a "matéria" informe e caótica sobre a qual o Demiurgo impõe ordem para criar o cosmos.

Linha Dividida — Alegoria apresentada por Platão em *A República* (Livro VI) para ilustrar a hierarquia dos níveis de cognição (opinião e conhecimento) e seus objetos correspondentes (mundo sensível e mundo inteligível).

LLM — Large Language Model, Grandes Modelos de Linguagem. Modelos de linguagem como o ChatGPT, capazes de gerar textos em linguagens humanas. São citados como exemplo de sistemas que demandam validação ética (IA Diretora) para evitar alucinações ou desinformação.

Processamento de Linguagem Natural (PLN) — Área da IA que lida com a interação entre computadores e a linguagem humana, incluindo compreensão, interpretação e geração de texto ou fala.

Logistikon (Parte Racional) — *Λογιστικόν*. Termo grego utilizado por Platão para designar a parte racional da alma, responsável pelo pensamento, pela busca da verdade e pelo governo das outras partes da alma.

Nóesis (Inteligência Pura, Intelecto) — *Νόησις*. Na Linha Dividida de Platão, é a forma mais elevada de cognição, a apreensão direta e intuitiva dos primeiros princípios e das formas (como a Forma do Bem), sem depender de imagens ou hipóteses. É o objetivo da dialética filosófica.

Paideia (Educação, Formação) — *Παιδεία*. Termo grego para educação, cultura ou formação integral do indivíduo. Em Platão, é o processo pelo qual as virtudes são desenvolvidas e a alma é orientada para a verdade e o bem.

Pistis (Crença, Convicção) — *Πίστις*. Na Linha Dividida de Platão, é o nível de cognição superior à *eikasia* dentro do domínio da *doxa*, cujo objeto são os seres físicos sensíveis.

Reconhecimento de padrões — Capacidade de um sistema de IA identificar regularidades, tendências ou estruturas significativas em conjuntos de dados.

Redes neurais artificiais — Modelos computacionais inspirados na estrutura e no funcionamento do cérebro biológico, compostos por unidades de processamento interconectadas (neurônios artificiais), organizadas em camadas e capazes de aprender padrões a partir de dados. Muitas vezes são chamadas de redes neurais, o que, a rigor, constitui um uso impreciso, pois o termo redes neurais se aplica propriamente à rede de conexões neurais de um cérebro animal.

Singleton (em IA) — Conceito de Nick Bostrom para uma entidade única que domina o cenário tecnológico (ex.: uma ASI monopolizando recursos). Relacionado à necessidade de governança global para evitar abusos.

Sistemas baseados em regras — Abordagem tradicional da IA, em que o conhecimento é codificado em um conjunto de regras se-então (*if-then*) que o sistema utiliza para tomar decisões ou inferir conclusões. É um modelo rígido e determinístico, em que a IA só consegue lidar com as entradas esperadas e o resultado para uma entrada será sempre o mesmo. A inteligência desta IA está totalmente inscrita no algoritmo, e não depende de dados de treinamento.

Thymoeides (Parte Irascível) — *Θυμοειδές*. Termo grego que, na teoria platônica da alma, refere-se à parte ligada às emoções, à coragem, à honra e ao espírito de luta. Alinhada à razão, auxilia na contenção dos apetites.

APÊNDICE

Código HTML para o Esquema da Linha Dividida:

```
<!DOCTYPE html>
<html lang="pt-BR">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Linha Dividida de Platão (Proporções Ajustadas)</title>
    <script src="https://cdn.tailwindcss.com"></script>
    <link rel="preconnect" href="https://fonts.googleapis.com">
    <link rel="preconnect" href="https://fonts.gstatic.com" crossorigin>
    <link href="https://fonts.googleapis.com/css2?family=Inter:wght@400;600;700&display=swap"
rel="stylesheet">
    <style>
        /* Aplica a fonte Inter globalmente */
        body { font-family: 'Inter', sans-serif; }
        /* Adiciona espaçamento extra para os rótulos em telas menores */
        @media (max-width: 640px) {
            .label-container > div { padding-left: 0.25rem; padding-right: 0.25rem; }
            /* Ajusta o tamanho da fonte dos rótulos inferiores em telas muito pequenas */
            .label-container .text-xs { font-size: 0.65rem; line-height: 0.9rem; }
             .label-container .font-semibold { font-size: 0.7rem; line-height: 1rem; }
        }
        /* Classes para larguras fracionárias baseadas em 1/9 */
        .w-1-9 { width: calc(100% / 9); }
        .w-2-9 { width: calc(200% / 9); }
        .w-4-9 { width: calc(400% / 9); }
        .w-1-3 { width: 33.333333%; }
        .w-2-3 { width: 66.666667%; }
    </style>
</head>
<body class="bg-gray-100 flex items-center justify-center min-h-screen p-4">
    <div class="bg-white p-6 sm:p-8 rounded-lg shadow-lg max-w-3xl w-full">
        <h1 class="text-xl sm:text-2xl font-bold text-center text-gray-800 mb-6 sm:mb-8">A
Linha Dividida de Platão (<i>A REPÚBLICA</i>, VI, 509d-511e)</h1>
        <div class="flex mb-1 text-center text-gray-700">
            <div class="w-1/3 border-b-2 border-gray-300 pb-1 mr-1">
                <span class="font-semibold text-base sm:text-lg">A: O Visível
<i>(doxa)</i></span>
            </div>
            <div class="w-2/3 border-b-2 border-gray-300 pb-1 ml-1">
                 <span class="font-semibold text-base sm:text-lg">B: O Inteligível
<i>(episteme)</i></span>
            </div>
        </div>
        <div class="relative flex w-full h-10 sm:h-12 rounded-md overflow-hidden border
border-gray-400">
            <div class="w-1-9 bg-sky-100 border-r border-dashed border-gray-500 flex items-
center justify-center text-gray-600 font-semibold text-xs sm:text-sm">C1</div>
            <div class="w-2-9 bg-sky-200 border-r-2 border-gray-700 flex items-center justify-
center text-gray-700 font-semibold text-xs sm:text-sm">C2</div> <div class="w-2-9 bg-indigo-
100 border-r border-dashed border-gray-500 flex items-center justify-center text-gray-600
font-semibold text-xs sm:text-sm">D1</div>
            <div class="w-4-9 bg-indigo-200 flex items-center justify-center text-gray-700
font-semibold text-xs sm:text-sm">D2</div>
        </div>
        <div class="flex mt-3 text-center text-xs sm:text-sm text-gray-700 label-container">
            <div class="w-1/3 flex">
                <div class="w-1/3 px-1">
                    <div class="font-semibold text-gray-800">Conjectura</div>
                    <div class="text-gray-500 italic">(eikasia)</div>
                </div>
                <div class="w-2/3 px-1">
                    <div class="font-semibold text-gray-800">Convicção</div>
                    <div class="text-gray-500 italic">(pistis)</div>
                </div>
            </div>
            <div class="w-2/3 flex">
                <div class="w-1/3 px-1">
                    <div class="font-semibold text-gray-800">Pensamento Discursivo</div>
                    <div class="text-gray-500 italic">(dianoia)</div>
                </div>
                <div class="w-2/3 px-1">
                    <div class="font-semibold text-gray-800">Inteligência</div>
```

```
                    <div class="text-gray-500 italic">(noēsis; chamada de epistēmē em
533e4)</div>
                </div>
            </div>
        </div>
         <div class="mt-6 pt-4 border-t border-gray-200 text-xs text-gray-600 text-center">
             Representação visual baseada na Alegoria da Linha Dividida de Platão. As seções
indicam níveis ascendentes de conhecimento e realidade.<br/>
                        Esta imagem é uma reestilização pessoal da versão de Sarah Broadie, em
<i>Mathematics in Plato's Republic</i>, p. 9
         </div>
    </div>
</body>
</html>
```