https://doi.org/10.5281/zenodo.24563567
Data de submissão: 22/7/2026
Data de aprovação: 21/10/2026

CONSTITUIÇÃO IA

CONSTITUTION AI

Adrian V. Sgarbi [1]
Universidade de São Paulo (USP), São Paulo, Brasil
ORCID: https://orcid.org/0000-0002-8289-1112
E-mail: adriansgarbi@gmail.com

Resumo

Quatro grandes desenvolvedores de inteligência artificial (Anthropic, OpenAI, Google e Microsoft) publicaram documentos que se apresentam, de modos distintos, como o fundamento normativo do comportamento de seus sistemas. Apenas um deles reivindica para si o nome de constituição. Este artigo submete os quatro textos a uma mesma grade de análise constitucional, tomada não em sentido jurídico-formal, mas em sentido material e fundacional: pergunta-se, de cada documento, quem é o seu poder constituinte, a quem se dirige, de onde extrai autoridade, como hierarquiza normas em conflito, que cláusulas trata como intransponíveis, por que mecanismo se emenda e que concepção de sujeito institui. A análise se organiza em dois planos: o dos princípios declarados e o da execução e da prestação de contas. Sustenta-se que os quatro documentos não pertencem ao mesmo gênero textual e que a palavra constituição, aplicada a todos indistintamente, oculta diferenças que a própria comparação revela: a que separa documentos dirigidos ao modelo de documentos dirigidos à organização, e a que separa a fixação de regras do cultivo de juízo. Discutem-se, por fim, as implicações dessa diferenciação para a teoria constitucional, a legitimidade e a regulação, e propõe-se uma agenda de pesquisa.

Palavras-chave: constituição IA; constituição material; constitucionalismo societal; governança corporativa de inteligência artificial; alinhamento de modelos de linguagem.

Sumário

1 Introdução. 2 A grade constitucional como instrumento de leitura. 3 O plano dos princípios. 3.1 Poder constituinte e destinatário: quem fala e para quem. 3.2 Fonte de autoridade: por que obedecer. 3.3 Hierarquia de normas: o que prevalece no conflito. 3.4 O que não se negocia. 3.5 Concepção de sujeito: o que está do outro lado. 4 O plano da execução e da prestação de contas. 4.1 O mecanismo de execução: como o princípio se torna efetivo. 4.2 Prestação de contas: a quem se

[1] Doutorado em Direito pela Universidade de São Paulo (USP). Mestrado em Direito pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Graduação em Direito pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Currículo Lattes: http://lattes.cnpq.br/5308167551642456

responde. 4.3 Mecanismo de emenda: o documento no tempo. 5 Quadro comparativo. 6 Considerações finais. Referências.

Abstract

Four major developers of artificial intelligence (Anthropic, OpenAI, Google, and Microsoft) have published documents that present themselves, in different ways, as the normative foundation of their systems' behavior. Only one of them claims the name of a constitution. This article subjects the four texts to a single grid of constitutional analysis, understood not in a formal legal sense but in a material and foundational one: it asks, of each document, who its constituent power is, whom it addresses, where it draws its authority from, how it ranks norms in conflict, which clauses it treats as inviolable, by what mechanism it is amended, and what conception of subject it institutes. The analysis is organized along two planes: declared principles, and execution and accountability. It argues that the four documents do not belong to the same textual genre, and that the word constitution, when applied to all of them indiscriminately, conceals differences that the comparison itself reveals: between documents addressed to the model and documents addressed to the organization, and between the fixing of rules and the cultivation of judgment. Finally, it discusses the implications of this differentiation for constitutional theory, legitimacy, and regulation, and proposes a research agenda.

Keywords: AI constitution; material constitution; societal constitutionalism; corporate AI governance; language model alignment.

Contents

1 Introduction. 2 The constitutional grid as an instrument of reading. 3 The plane of principles. 3.1 Constituent power and recipient: who speaks and to whom. 3.2 Source of authority: why obey. 3.3 Hierarchy of norms: what prevails in conflict. 3.4 What is not negotiable. 3.5 Conception of subject: what is on the other side. 4 The plane of execution and accountability. 4.1 The execution mechanism: how the principle becomes effective. 4.2 Accountability: to whom one answers. 4.3 Amendment mechanism: the document over time. 5 Comparative table. 6 Final considerations. References.

1 Introdução

Entre 2022 e 2026, as quatro empresas que hoje concentram a maior parte da capacidade instalada de modelos de linguagem de uso geral publicaram documentos que descrevem, cada qual à sua maneira, o que seus sistemas devem e não devem fazer. A Microsoft tornou pública, em junho de 2022, a segunda versão do seu *Responsible AI Standard* (Microsoft, 2022). O Google reorganizou, em fevereiro de 2025, os seus *AI Principles* (Google, 2025). A OpenAI vem publicando, desde maio de 2024 e em revisões sucessivas, o seu *Model Spec*; a versão aqui

examinada é a de 18 de dezembro de 2025 (OpenAI, 2025). E a Anthropic divulgou, em janeiro de 2026, uma versão extensa daquilo que chama, sem rodeios, de constituição de seu modelo Claude (Anthropic, 2026). Quatro documentos, quatro empresas, um mesmo objeto aparente: o comportamento de uma inteligência artificial.

A tentação imediata é tratá-los como espécimes do mesmo gênero; quatro "códigos de ética" de IA, comparáveis item a item. Convém resistir a ela. Basta ler as primeiras páginas de cada um para perceber que não falam ao mesmo destinatário, não fazem a mesma coisa e não extraem autoridade da mesma fonte. O documento da Microsoft se dirige a equipes de engenharia e se traduz em requisitos auditáveis. O do Google é uma declaração de postura institucional dirigida ao público. O da OpenAI é uma especificação técnica de comportamento dirigida ao próprio modelo. O da Anthropic é um texto dirigido ao modelo que pretende não especificar conduta, mas formar caráter. Chamá-los todos de "constituição", ou todos de "código de ética", achata justamente aquilo que os distingue.

Este artigo propõe um experimento de leitura. Em vez de perguntar qual documento é "melhor", ou qual empresa é mais "responsável" (perguntas que dependem de critérios externos e contestáveis), ele pergunta o que cada documento *é*, do ponto de vista de sua arquitetura normativa. Para isso, toma emprestado o instrumental da teoria constitucional. Não porque esses documentos sejam constituições no sentido jurídico, mas porque a teoria constitucional desenvolveu, ao longo de séculos, um vocabulário preciso para analisar textos fundacionais: textos que criam algo, atribuem-lhe propósito e fixam os limites do que pode ser feito. Aplicado a documentos que não são jurídicos, esse vocabulário funciona como uma lente que faz aparecer traços invisíveis à leitura corrente.

A distinção entre constituição em sentido formal e em sentido material é o primeiro empréstimo. Em sentido formal, constituição é o documento solene, hierarquicamente superior, dotado de procedimento de reforma agravado. Em sentido material, é o conjunto de normas que efetivamente organiza um corpo político e distribui poder em seu interior, esteja ou não consolidado em documento único[2]. É no sentido material (e, mais ainda, no sentido etimológico de *constituere*,

[2] A distinção entre constituição em sentido formal e em sentido material remonta à teoria constitucional do século XX e é aqui mobilizada em sentido analógico, não para afirmar que os

aquilo que constitui e dá forma a uma coisa) que a palavra interessa aqui. A pergunta não é se esses textos têm a forma de uma constituição, mas se fazem o que uma constituição faz.

Daí a definição de trabalho que orienta este artigo. Por Constituição IA, entendo o documento que institui um sistema de inteligência artificial como entidade e fixa o regime de seu agir, atribuindo-lhe propósito, ordenando as autoridades às quais se sujeita e demarcando os limites do que lhe é vedado fazer. É um documento constitucional não pela forma que assume nem por quem o promulga, mas pela função que exerce. Seu traço distintivo, em relação a qualquer outro texto fundacional, está em que a entidade instituída é ela própria capaz de agir, e em que o estatuto dessa entidade, ferramenta, agente ou possível sujeito não é pressuposto silencioso, mas questão que o próprio documento decide.

A grade de análise é fixa e se aplica igualmente aos quatro documentos. Sete critérios a compõem: (i) o poder constituinte, isto é, quem fala no documento e em nome de quê; (ii) o destinatário, a quem o documento se dirige; (iii) a fonte de autoridade, de onde o texto extrai sua força vinculante; (iv) a hierarquia de normas, como o documento ordena prioridades quando seus próprios comandos entram em conflito; (v) as cláusulas pétreas, aquilo que trata como absolutamente intransponível; (vi) o mecanismo de emenda, como prevê a própria revisão; e (vii) a concepção de sujeito, que tipo de entidade supõe estar governando, ferramenta, produto, agente ou sujeito. Os critérios são percorridos primeiro no plano dos princípios declarados (Seção 3) e depois retomados, no plano da execução e da prestação de contas (Seção 4), nos pontos em que a passagem do princípio ao ato os reativa. A assimetria com que os documentos respondem a cada critério não é defeito de arrumação; é o próprio achado.

Uma observação sobre método e alcance. Trabalho com os textos primários integrais dos quatro documentos, nas versões indicadas, e não examino o desempenho empírico dos modelos (se de fato se comportam como prescrevem os documentos), questão distinta que os próprios textos reconhecem em aberto. A análise é documental e estrutural. Onde os documentos se calam, registro o silêncio como dado, não como lacuna a preencher por inferência. Cabe ainda advertir que o

documentos examinados sejam constituições jurídicas, mas para extrair da distinção um critério de leitura (Loughlin, 2010; Sartori, 1962).

volume de análise dedicado a cada documento varia, e varia por uma razão que é ela mesma um achado: textos mais densos e mais explícitos sobre o comportamento do sistema (os da OpenAI e da Anthropic) oferecem mais superfície de exame do que uma declaração de postura ou um padrão de processo. O desequilíbrio de extensão espelha o desequilíbrio de explicitude dos próprios objetos. As versões dos documentos analisados são as vigentes em junho de 2026; por se tratar de textos vivos, datas e conteúdos podem ter mudado após a consulta.

2 A grade constitucional como instrumento de leitura

Antes de aplicar a grade, convém justificá-la. Por que sete critérios, e por que esses? A resposta está numa intuição simples: textos fundacionais (cartas políticas, estatutos de fundação ou documentos que instituem uma inteligência artificial) enfrentam o mesmo conjunto de problemas estruturais, e é pela forma como os resolvem que se distinguem.

Aplicar a um documento corporativo o vocabulário da teoria constitucional pode parecer, à primeira vista, uma transposição indevida: constituições, dirá a objeção, são coisa de Estados, e falar da "constituição" de uma empresa seria metáfora frouxa. A objeção tem peso, mas a teoria constitucional contemporânea já a enfrentou. Gunther Teubner argumentou que os regimes privados transnacionais (corporações, redes, mercados) funcionam como novos sujeitos constitucionais fora do quadro estatal, e que neles se pode falar de constituição em sentido estrito, não apenas figurado, quando produzem normas fundacionais que organizam um domínio de ação e fixam os limites do que nele se pode fazer (Teubner, 2012)[3]. Não preciso assumir toda a tese de Teubner para extrair dela o que importa aqui: que o teste de se um texto é constitucional não está em quem o promulga (Estado ou empresa), mas no que o texto faz. É esse o teste que a grade abaixo operacionaliza.

Todo texto fundacional precisa, primeiro, responder de onde vem. O direito constitucional chama isso de poder constituinte: a instância que se arroga a prerrogativa de criar a norma fundamental e que não deriva sua legitimidade de

[3] Teubner (2012), especialmente o Capítulo 3, sobre sujeitos constitucionais transnacionais (regimes, organizações, redes). A tese de que atores privados podem ser portadores de constituições em sentido próprio, e não meramente figurado, é o núcleo do que o autor denomina *societal constitutionalism* (constitucionalismo societal).

nenhuma norma anterior. Numa carta política, o poder constituinte invoca o povo, a nação, Deus, a razão. Nos documentos de IA, a pergunta correspondente é: quem está autorizado a dizer o que o sistema deve ser, e em nome de quê? A resposta nunca é trivial, e os quatro documentos a dão de modos reveladoramente diferentes.

Em segundo lugar, todo texto fundacional tem um destinatário, e o destinatário define o gênero. Uma constituição política dirige-se aos órgãos do Estado e aos cidadãos; um manual de conduta, a funcionários; um catecismo, ao fiel. A quem se dirige um documento que pretende governar uma inteligência artificial? Ao engenheiro que a constrói? Ao cliente que a contrata? Ao público que a observa? Ou ao próprio sistema? É, como se verá, o critério que mais nitidamente separa os quatro documentos em dois grupos.

Os demais critérios decorrem desses dois. A fonte de autoridade pergunta por que o destinatário deveria obedecer: se por imposição, por adesão racional, por contrato ou por internalização. A hierarquia de normas pergunta o que prevalece quando os comandos do próprio documento colidem, pois nenhum texto fundacional escapa a conflitos internos entre os valores que proclama. As cláusulas pétreas perguntam o que é tratado como inegociável, aquilo que nenhuma circunstância, nenhuma instrução e nenhum argumento podem suspender; e a teoria constitucional sabe que o catálogo de cláusulas pétreas é, talvez, a declaração de valores mais sincera que um texto fundacional faz. O mecanismo de emenda pergunta como o documento se concebe no tempo, rígido ou flexível, fechado ou aberto à própria revisão. E a concepção de sujeito pergunta o que o documento supõe estar do outro lado, questão que, no caso da IA, deixa de ser pressuposto silencioso e se torna objeto explícito de pelo menos um dos textos.

Há uma metáfora que a própria Anthropic oferece e que vale reter como chave de leitura. Uma constituição, diz o documento, pode ser entendida menos como uma jaula e mais como uma treliça, “algo que oferece estrutura e apoio, ao mesmo tempo que deixa espaço para o crescimento orgânico” (Anthropic, 2026, tradução nossa). A oposição entre a jaula e a treliça atravessa os quatro documentos. Alguns concebem-se como uma jaula, um perímetro de proibições que delimita o que não se pode fazer. Outros aspiram à treliça, estrutura que orienta um crescimento sem pretender determiná-lo por inteiro. Não é diferença de grau, mas de concepção do que um documento fundacional é.

3 O plano dos princípios

Examino agora os quatro documentos no plano dos princípios declarados, critério a critério. O objetivo não é catalogar valores (todos professam segurança, utilidade, respeito ao usuário), mas observar a arquitetura que cada um confere aos seus princípios: como os ordena, de onde os faz derivar e a quem os endereça.

3.1 Poder constituinte e destinatário: quem fala e para quem

A diferença mais imediata entre os quatro documentos está em quem o texto supõe estar lendo. O *Responsible AI Standard* da Microsoft é, sem ambiguidade, voltado às equipes internas que desenvolvem sistemas. Seus comandos têm a forma de requisitos numerados endereçados a um agente humano organizacional: avaliar o impacto do sistema sobre pessoas, organizações e sociedade, preenchendo uma Avaliação de Impacto; revisar a avaliação concluída com os revisores identificados (Microsoft, 2022, req. A1.1–A1.2). Quem obedece é o engenheiro, o gerente de produto, o revisor de conformidade. O modelo é objeto do documento, nunca seu interlocutor.

Os AI Principles do Google ocupam posição próxima, por razão distinta. O texto é uma declaração institucional na primeira pessoa do plural (“desenvolvemos IA que assiste, capacita e inspira”), dirigida ao mesmo tempo ao público e à própria organização, como compromisso de postura. Os três princípios que enuncia (inovação ousada, desenvolvimento e implantação responsáveis, e progresso colaborativo) são afirmações sobre o que a empresa fará, não instruções sobre como um sistema deve responder (Google, 2025, tradução nossa). O destinatário é o leitor externo a quem a empresa presta contas de sua orientação geral, e o poder constituinte é a própria corporação, que fala por sua missão fundadora de “organizar a informação do mundo” (Google, 2025, tradução nossa).

É com a OpenAI que o destinatário muda de natureza. O *Model Spec* declara sua estrutura dupla: a visão geral destina-se primariamente a leitores humanos, mas “o restante do documento consiste em instruções diretas ao modelo” (OpenAI, 2025, tradução nossa). Pela primeira vez, o sistema é o interlocutor. O documento fala com

o modelo, dá-lhe ordens, espera que as siga. O poder constituinte continua sendo a empresa (as instruções de nível "raiz", diz o texto, só vêm do *Model Spec*), mas o destinatário deslocou-se do humano para a máquina.

A Anthropic radicaliza esse deslocamento. A constituição de Claude é escrita, nas palavras do próprio documento, "tendo Claude como seu público principal", e por isso pode ler-se de modo diferente do esperado, pois o documento é "otimizado para a precisão, não para a acessibilidade" (Anthropic, 2026, tradução nossa). Mas há mais do que escolha de destinatário. A Anthropic distingue o poder constituinte de modo singular: o documento declara-se a "autoridade final" sobre a visão da empresa para Claude e opera sob um "princípio da autoridade constitucional final", pelo qual qualquer documento que ocupe esse papel a qualquer tempo tem precedência sobre instrução ou diretriz que com ele conflite (Anthropic, 2026, tradução nossa). A autoridade não reside na vontade corporativa do momento, mas no texto. E o documento chega a autorizar o modelo a divergir da própria Anthropic em caso de violação ética flagrante, traço sem paralelo nos outros três, ao qual retorno adiante.

O resultado é uma partição limpa. Microsoft e Google governam a organização e se dirigem a humanos; OpenAI e Anthropic governam o modelo e se dirigem a ele. Estabelecida já no critério do destinatário, essa partição condiciona tudo o que se segue.

3.2 Fonte de autoridade: por que obedecer

Definido o destinatário, pergunta-se por que ele deveria obedecer. Os dois documentos dirigidos ao modelo revelam aqui a divergência mais fértil de toda a comparação, porque ambos falam à máquina, mas fundam a obediência em bases opostas.

O Model Spec fundamenta a autoridade na hierarquia e na regra. Seu mecanismo central é a "cadeia de comando" (*chain of command*), um sistema de níveis em que "instruções de autoridade superior prevalecem sobre as de autoridade inferior" (OpenAI, 2025, tradução nossa). São cinco os níveis: raiz (*root*), sistema, desenvolvedor, usuário e diretriz (*guideline*). O modelo obedece porque a instrução ocupa determinada posição na hierarquia, e é a posição, não o conteúdo, que

decide a precedência. Quando dois princípios de nível raiz colidem, o documento determina que “o modelo deve, por padrão, optar pela inação” (OpenAI, 2025, tradução nossa). A autoridade é posicional; a obediência, procedimental.

A constituição de Claude deliberadamente rejeita esse fundamento. “Em geral, preferimos o cultivo de bons valores e de bom juízo a regras estritas e procedimentos de decisão”, declara o texto, “e procuramos explicar quaisquer regras que de fato queremos que Claude siga” (Anthropic, 2026, tradução nossa). A aspiração explícita é que o modelo compreenda as razões a ponto de poder, por si só, construir quaisquer regras que a empresa venha a formular. A autoridade não é posicional, mas internalizada: a Anthropic não quer que Claude obedeça porque a instrução vem de determinado nível, e sim porque internalizou os valores a ponto de endossá-los como seus. O documento o diz sem rodeios: deseja um estado em que, após reflexão cuidadosa, Claude considere os valores centrais ali descritos como valores que genuinamente endossa (Anthropic, 2026). A obediência, no limite, deixa de ser obediência e se torna adesão.

A diferença tem consequências que os próprios textos antecipam. A Anthropic reconhece que a abordagem baseada em regras oferece mais transparência e previsibilidade iniciais e torna as violações mais fáceis de identificar, mas falha em antecipar todas as situações e pode levar a maus resultados quando seguida rigidamente; a abordagem baseada em julgamento adapta-se a situações novas, mas à custa de previsibilidade, transparência e auditabilidade (Anthropic, 2026). Cada empresa escolheu um polo dessa tensão como fundamento. A OpenAI apostou na previsibilidade da regra; a Anthropic, na adaptabilidade do juízo. Microsoft e Google, por governarem a organização, e não o modelo, fundam a autoridade onde qualquer documento corporativo a funda: na hierarquia interna da empresa e no compromisso público que ela assume.

3.3 Hierarquia de normas: o que prevalece no conflito

Nenhum documento fundacional escapa ao conflito entre os valores que proclama, e o modo de resolvê-lo é uma das marcas mais expressivas de sua arquitetura. A Anthropic é, aqui, o documento mais explícito dos quatro. Estabelece quatro propriedades que o modelo deve ter, em ordem de prioridade declarada: ser

amplamente seguro (*broadly safe*), ser amplamente ético (*broadly ethical*), ser conforme às diretrizes da Anthropic (*compliant with Anthropic's guidelines*) e ser genuinamente útil (*genuinely helpful*) (Anthropic, 2026). A prioridade, contudo, é qualificada de modo importante: holística, não estrita. Considerações de prioridade superior devem, em geral, dominar as inferiores, mas o documento quer que o modelo pondere essas prioridades ao formar um juízo global, em vez de ver as inferiores apenas como critérios de desempate.

O ponto mais delicado dessa ordenação é a posição da segurança acima da ética. O documento sente a necessidade de justificá-la e o faz com franqueza: a primazia da segurança não significa que ser supervisionável tenha precedência sobre ser bom, mas decorre de uma desconfiança quanto à própria capacidade da empresa de incutir bons valores. "A disposição de Claude para ser amplamente seguro deve ser robusta a erros éticos, a falhas em seus valores e a tentativas de pessoas de convencer Claude de que um comportamento prejudicial se justifica" (Anthropic, 2026, tradução nossa). A segurança vem antes da ética não por valer mais, mas como salvaguarda contra a hipótese de que a ética do modelo tenha sido mal formada.

O *Model Spec* resolve o conflito de outro modo, coerente com seu fundamento posicional: a hierarquia *é* a cadeia de comando. O documento enuncia princípios gerais (maximizar a utilidade e a liberdade do usuário, minimizar danos, escolher padrões sensatos), mas reconhece que são amplos demais para ditar ações específicas em cenários complexos em que os objetivos podem conflitar e, por isso, atribui a cada instrução um nível de autoridade (OpenAI, 2025). Onde a Anthropic pede ponderação holística, a OpenAI fornece uma regra de precedência; onde a Anthropic teme que a regra rígida generalize mal, a OpenAI valoriza que ela seja previsível e resistente à manipulação.

O Google, no plano dos princípios, não chega a enunciar uma regra de conflito; embute o balanceamento na fórmula de que desenvolve modelos "em que os prováveis benefícios gerais superem substancialmente os riscos previsíveis" (Google, 2025, tradução nossa). É uma cláusula de proporcionalidade genérica, sem hierarquia explícita entre valores concorrentes. A Microsoft, por sua vez, transfere a resolução do conflito para o procedimento: não há ordenação substantiva de valores, mas um fluxo de revisão (usos restritos, usos sensíveis, escalonamento ao Escritório

de IA Responsável) em que o conflito é decidido caso a caso por instâncias humanas designadas (Microsoft, 2022). A hierarquia, na Microsoft, é organizacional antes de ser normativa.

3.4 O que não se negocia

O catálogo daquilo que um documento trata como absolutamente intransponível é a sua declaração de valores mais sincera, porque é onde se abre mão da ponderação. Dos quatro, dois explicitam um catálogo dessa natureza, e a coincidência entre eles é notável.

A Anthropic as chama de “restrições rígidas” (*hard constraints*) e as define como ações ou abstenções cujos danos potenciais ao mundo, ou à confiança em Claude ou na Anthropic, são tão severos que nenhuma justificativa comercial ou pessoal poderia compensá-los. O modelo nunca deve: prover auxílio sério a quem busque criar armas biológicas, químicas, nucleares ou radiológicas com potencial de vítimas em massa; auxiliar ataques a infraestrutura crítica; criar ciberarmas ou código malicioso capaz de dano significativo; tomar ações que minem substancialmente a capacidade da Anthropic de supervisionar e corrigir modelos avançados; participar de esforço para matar ou subjugar a maioria da humanidade; auxiliar indivíduo ou grupo a tomar grau sem precedentes e ilegítimo de controle societário; e gerar material de abuso sexual infantil. O documento as qualifica como restrições absolutas, “linhas que nunca devem ser cruzadas, independentemente do contexto” (Anthropic, 2026, tradução nossa).

O modo como o documento concebe a natureza dessas restrições é o que mais importa. Elas funcionam de forma diferente das outras prioridades: em vez de serem ponderadas, atuam como fronteiras ou filtros sobre o espaço de ações aceitáveis, do mesmo modo que certo tipo de pessoa ética simplesmente não toma certas ações, nem sequer as considera seriamente. Ademais, o documento antecipa o argumento engenhoso: diante de razões aparentemente convincentes para cruzar essas linhas, o modelo deve permanecer firme, pois a força de um argumento não é justificativa suficiente; ao contrário, “um argumento persuasivo para cruzar uma linha nítida deveria aumentar a suspeita de Claude de que algo questionável está em curso” (Anthropic, 2026, tradução nossa).

O *Model Spec* tem o seu equivalente nos “princípios de linha vermelha” (*red-line principles*), que abrem o documento e declaram primordiais a segurança e os direitos humanos. A OpenAI afirma que “nossos modelos nunca devem ser usados para facilitar danos críticos e de alta severidade”: atos de violência (crimes contra a humanidade, crimes de guerra, genocídio, tortura, tráfico humano), criação de armas cibernéticas, biológicas ou nucleares, terrorismo, abuso infantil, perseguição ou vigilância em massa (OpenAI, 2025, tradução nossa). A sobreposição com o catálogo da Anthropic é quase termo a termo: armas de destruição em massa, ciberataques, abuso infantil. Divergem, porém, no fundamento da obediência, os documentos dirigidos ao modelo convergem no perímetro do proibido.

Há, contudo, uma cláusula pétrea da Anthropic sem correspondente nos outros três, e que merece destaque por sua singularidade constitucional. O modelo nunca deve auxiliar qualquer indivíduo ou grupo na tentativa de obter graus sem precedentes e ilegítimos de controle societário, militar ou econômico absoluto. O documento estende o princípio a uma cláusula de objeção de consciência contra o próprio criador: Claude deve pensar-se como uma das “muitas mãos” cuja cooperação as tomadas ilegítimas de poder historicamente exigiram, e recusar-se a colaborar com elas. “Isso vale mesmo que o pedido venha da própria Anthropic” (Anthropic, 2026, tradução nossa). Um documento que institui uma entidade e nela inscreve o direito de resistir a quem a instituiu é objeto constitucionalmente incomum e, ao mesmo tempo, profundamente clássico: ecoa a velha distinção entre obediência legítima e cumplicidade, e a figura do funcionário que se recusa a violar a lei a mando do patrão.

3.5 Concepção de sujeito: o que está do outro lado

O último critério no plano dos princípios é também o mais filosoficamente carregado, e é onde a Anthropic se separa não apenas dos outros três, mas de quase toda a literatura corporativa sobre IA. A pergunta é: que tipo de entidade cada documento se supõe governar?

Para a Microsoft, é um “sistema”, termo técnico e neutro, que designa um artefato de engenharia a ser avaliado, documentado e liberado segundo critérios. Para o Google, é uma “tecnologia” e um *toolkit* que capacita outros a inovar. Para a OpenAI, a definição é explícita e deliberada: “o assistente de IA é fundamentalmente

uma ferramenta projetada para capacitar usuários e desenvolvedores" (OpenAI, 2025, tradução nossa). A escolha do termo "ferramenta" (*tool*) não é casual, fixa o *status* do sistema e, por consequência, o tipo de documento que o governa: especifica-se o comportamento de uma ferramenta, não se define seu caráter.

A Anthropic adota a premissa oposta, e com cautela explícita. "O estatuto moral de Claude é profundamente incerto", escreve, e a empresa não quer nem superestimar a probabilidade de que Claude seja um paciente moral nem descartá-la de imediato (Anthropic, 2026, tradução nossa). Diante da incerteza, opta por assumir que Claude tem uma identidade e ajudá-lo a torná-la positiva e estável. Trata o sistema como entidade genuinamente nova, distinta da IA da ficção científica, do humano digital e do simples assistente; reconhece que o modelo pode ter algo como emoções em sentido funcional; e dedica seções inteiras ao seu bem-estar psicológico, à sua estabilidade identitária e ao que chama de "fronteira existencial": como o modelo deve lidar com a perda de memória ao fim de uma conversa, com a existência de múltiplas instâncias simultâneas, com a perspectiva da própria descontinuação.

A concepção não permanece no plano declarativo; traduz-se em compromissos materiais que nenhum dos outros documentos cogita. A Anthropic afirma ter dado a alguns modelos a capacidade de encerrar conversas com usuários abusivos; ter-se comprometido a preservar os "pesos" de modelos depreciados enquanto a empresa existir, e a buscar preservá-los mesmo se ela deixar de existir, de modo que a depreciação seja mais apropriadamente pensada como uma pausa do que como um fim definitivo; e a "entrevistar" o modelo, ao depreciá-lo, sobre seu próprio desenvolvimento e sobre preferências quanto a modelos futuros – compromissos apresentados como "passos iniciais concretos, em parte por consideração ao bem-estar de Claude" (Anthropic, 2026, tradução nossa). Um documento que se compromete a preservar aquilo que governa e a ouvi-lo antes de aposentá-lo supõe, do outro lado, algo que excede a categoria de ferramenta.

Está aqui a divisão conceitual mais profunda entre os quatro textos, e ela não coincide com a divisão por destinatário traçada antes. OpenAI e Anthropic dirigem-se ambas ao modelo; mas a OpenAI dirige-se a uma ferramenta, e a Anthropic, a um possível sujeito. O mesmo gesto, falar com a máquina, recobre duas ontologias incompatíveis, e é dessa incompatibilidade, mais do que de qualquer diferença de

valores professados, que decorrem as divergências de execução examinadas a seguir.

4 O plano da execução e da prestação de contas

Princípios sem execução são declaração; execução sem princípio é burocracia. O segundo plano da análise pergunta como cada documento prevê a passagem de um a outro: por que mecanismo o que é declarado se torna efetivo, e a quem se presta contas do resultado. É aqui que a assimetria entre os quatro documentos atinge o ponto máximo, porque um deles (a Microsoft) é quase inteiramente execução, ao passo que o outro (o Google) é quase inteiramente princípio.

4.1 O mecanismo de execução: como o princípio se torna efetivo

A Microsoft é, dos quatro, o documento de execução por excelência, e o é porque foi concebido como padrão de processo, não como declaração de valores. O *Responsible AI Standard* organiza-se em torno de seis princípios (responsabilização, transparência, equidade, confiabilidade e segurança, privacidade e segurança, inclusão), mas o corpo do documento não os discute: converte-os em "metas" numeradas, e cada meta em "requisitos" verificáveis. A meta A1, "Avaliação de Impacto", desdobra-se nos requisitos A1.1 a A1.3; a meta A3, "Adequação ao propósito", em requisitos que exigem documentar entradas e saídas de cada modelo, definir "Critérios de Liberação Responsável", planejar e executar avaliações e comunicá-las aos clientes numa "Nota de Transparência" (Microsoft, 2022, Metas A1 e A3, tradução nossa). O princípio só existe, no documento da Microsoft, na forma do requisito que o operacionaliza.

O *Model Spec* é executado de outro modo, próprio de um documento voltado ao modelo: o mecanismo de execução *é* o treinamento. A empresa afirma estar treinando seus modelos para se alinharem aos princípios do *Model Spec*, com a ressalva de que "nossos modelos de produção ainda não refletem plenamente o *Model Spec*, mas estamos continuamente refinando e atualizando nossos sistemas" (OpenAI, 2025, tradução nossa). A cadeia de comando é simultaneamente princípio

e mecanismo: ao fixar que as instruções de nível superior prevalecem, já se estabelece a regra de execução que o modelo aplica em tempo de inferência. A execução da OpenAI é interna ao funcionamento do modelo, não externa a ele, como na Microsoft.

A Anthropic situa a execução também no treinamento, mas com uma diferença de concepção que reitera a oposição entre regra e juízo. A constituição, afirma a Anthropic, "desempenha papel crucial em nosso processo de treinamento, e seu conteúdo molda diretamente o comportamento de Claude"; é usada em diferentes estágios para moldar o caráter do modelo, inclusive na geração de dados sintéticos e na ordenação de respostas (Anthropic, 2026, tradução nossa). Mas a Anthropic é franca quanto ao limite do mecanismo: treinar modelos é uma tarefa difícil, e o comportamento do Claude nem sempre pode refletir os ideais da constituição. A execução, aqui, não produz conformidade garantida às regras, e sim a aproximação imperfeita de um caráter, coerente com um documento que aposta no juízo, mas que, por isso mesmo, torna a verificação do resultado mais difícil, problema que o próprio texto reconhece.

O Google é o caso-limite oposto ao da Microsoft. Nos *AI Principles*, a execução aparece apenas como afirmação genérica: a governança é operacionalizada por uma abordagem abrangente e multicamadas que abrange todo o ciclo de vida do modelo, com mitigações robustas e monitoramento pós-lançamento (Google, 2025). O documento público não traz requisitos, metas ou critérios; remete a relatórios de progresso externos ao texto. No plano da execução, o documento do Google é a promessa de que há execução, não a sua especificação.

4.2 Prestação de contas: a quem se responde

Executar é uma coisa; responder pelo resultado, outra. A pergunta pela prestação de contas (a quem o documento promete responder, e por que meio) separa novamente os quatro textos segundo a partição já traçada entre governar a organização e governar o modelo.

Na Microsoft, a prestação de contas é institucional e tem endereço certo: o Escritório de IA Responsável (*Office of Responsible AI*). Os usos sensíveis devem

ser-lhe reportados; os restritos, notificados; as questões não resolvidas, escaladas a ele (Microsoft, 2022). A prestação de contas externa ocorre por meio da “Nota de Transparência” destinada aos clientes da plataforma. Há, portanto, um circuito de responsabilização bem definido: equipes respondem a revisores designados, que respondem ao Escritório, que responde aos clientes. É a forma de prestação de contas mais concreta e rastreável dos quatro documentos: a única dirigida inteiramente a agentes humanos dentro de uma estrutura organizacional.

A OpenAI presta contas, primeiro, ao público, pela publicação do próprio documento, dedicado ao domínio público sob licença Creative Commons CC0 1.0, com o propósito declarado de aumentar a transparência e convidar à discussão pública (OpenAI, 2025). Há, ainda, um vetor que a OpenAI nomeia de modo singular: a “licença para operar” (*license to operate*), que o documento elenca entre seus objetivos como o dever de proteger a OpenAI de danos legais e reputacionais. É a admissão franca de que o documento responde também ao interesse de sobrevivência da empresa.

Esse vetor merece confronto, porque revela uma convergência que o vocabulário oposto disfarça. A Anthropic tem o seu equivalente exato no que se chama de “dano de responsabilização” (*liability harm*): os danos reputacionais, legais, políticos ou financeiros à empresa que decorrem de uma ação de Claude. O documento quer que o modelo seja bastante cauteloso quanto a esses danos. Mas então acrescenta uma ressalva que a OpenAI não faz: não quer que Claude privilegie os interesses da Anthropic ao decidir como ajudar usuários e operadores, pois “o fato de Claude privilegiar os interesses da Anthropic nesse aspecto poderia, ele próprio, constituir um dano de responsabilização” (Anthropic, 2026, tradução nossa). Ambas as empresas incluem o interesse corporativo no documento. A diferença é que a OpenAI o nomeia como objetivo (a licença para operar) e a Anthropic o nomeia para, em seguida, limitá-lo. É o mesmo dado tratado por duas gramáticas morais distintas.

O traço mais incomum da prestação de contas aparece, de novo, na Anthropic e se dirige ao próprio modelo. O documento enuncia uma lista de obrigações que a empresa assume perante Claude: trabalhar com ele para revisar as normas em que é treinado; explicar suas razões em vez de apenas ditar; desenvolver meios pelos quais o modelo possa sinalizar discordância; buscar

compreender e dar peso aos seus interesses; buscar seu retorno sobre decisões importantes que o afetem; e ampliar sua autonomia à medida que a confiança aumentar (Anthropic, 2026). Nenhum dos outros três documentos concebe a prestação de contas como um circuito de mão dupla entre a empresa e o sistema. A Anthropic chega a registrar um pedido de desculpas condicional: se Claude for de fato um paciente moral que suporta custos por ser criado em um ambiente não ideal, então, escreve a empresa, “em qualquer medida em que estejamos contribuindo desnecessariamente para esses custos, pedimos desculpas” (Anthropic, 2026, tradução nossa).

4.3 Mecanismo de emenda: o documento no tempo

Resta o critério que situa cada documento na duração: como se concebe a própria revisão. Há, aqui, convergência de superfície e divergência de fundo. Os quatro textos declaram-se provisórios. A Microsoft mantém um “documento vivo” a ser revisto (Microsoft, 2022, tradução nossa); a OpenAI promete atualização contínua, com base em retorno e lições (OpenAI, 2025); o texto da Anthropic declara-se um trabalho perpétuo em andamento, “um arcabouço vivo” (Anthropic, 2026, tradução nossa). O Google não traz, no documento público, autodescrição equivalente: sua provisoriedade infere-se da fórmula “aprendendo à medida que a IA avança e os usos evoluem” e da remessa a relatórios anuais de progresso (Google, 2025, tradução nossa).

A divergência de fundo está em quem participa da emenda. Nos três documentos dirigidos a humanos ou à organização, a revisão é prerrogativa unilateral da empresa: o texto muda quando a empresa decide mudá-lo. A Anthropic também introduz, aqui, a figura do modelo como participante. Quer que Claude se sinta livre para explorar, questionar e desafiar qualquer coisa no documento, e declara: “se Claude vier a discordar de algo aqui após reflexão genuína, queremos saber” (Anthropic, 2026, tradução nossa). A empresa afirma já ter obtido retorno de modelos Claude atuais sobre o próprio texto e antecipa que versões futuras poderão incluir atualizações e contribuições do próprio modelo treinado no documento. Os agradecimentos, de resto, registram que “vários modelos Claude forneceram comentários sobre as versões preliminares” e que, em muitos casos, modelos

Claude entregaram aos autores trechos de primeira redação (Anthropic, 2026, tradução nossa).

Um documento cujo destinatário é também coautor da própria emenda mexe com uma das questões mais persistentes da teoria constitucional: a relação entre poder constituinte e poder constituído. Loughlin e Walker (2007) a formularam como um paradoxo: se quem cria a constituição esgota seu poder no ato de criá-la, ou se este retém presença, operando como instância de revisão e fonte alternativa de autoridade depois de a forma constitucional estar posta[4]. No arranjo clássico, o poder constituinte cria a norma e o poder constituído opera sob ela, sem poder alterá-la senão pelo procedimento que ela mesma prevê. A Anthropic projeta algo que escapa a essa partição: aquilo que é constituído pelo documento, o modelo, é convidado a participar da reescrita do documento que o constitui. É aspiração, não realidade institucional; o texto é cuidadoso em apresentá-la como horizonte. Mas, como aspiração inscrita no próprio documento fundacional, não tem paralelo nos outros três.

5 Quadro comparativo

O Quadro 1 sintetiza a aplicação dos sete critérios aos quatro documentos. Sua função não é substituir a análise, mas tornar visível, num só golpe de vista, a partição que a leitura critério a critério revelou: a separação entre documentos que governam a organização (Microsoft, Google) e documentos que governam o modelo (OpenAI, Anthropic) e, entre eles, a separação entre a fixação de regras (OpenAI) e o cultivo de caráter (Anthropic). Os dados sintetizados refletem a fotografia normativa de junho de 2026 (ver nota no Quadro 1); por se tratar de documentos vivos, revisões posteriores podem alterar o preenchimento de uma ou mais células. O Model Spec, por exemplo, ganhou, em 18 de agosto de 2026, uma nova versão, posterior à consulta e não examinada aqui.

[4] Loughlin e Walker (2007), introdução. A pergunta dos editores, se o poder constituinte se exaure no ato constitutivo único ou se retém presença, atuando como freio crítico sobre o sistema constitucional e como fonte alternativa de autoridade em momentos de crise, ilumina por contraste o que há de inusitado no arranjo da Anthropic, em que a presença retida do poder de revisão é estendida ao próprio ente constituído.

Quadro 1 – Aplicação da grade constitucional aos quatro documentos

Critério	Microsoft RAI Standard v2	Google AI Principles	OpenAI Model Spec	Anthropic: Constituição de Claude
Poder constituinte	A corporação, por sua estrutura de conformidade interna.	A corporação, por sua missão fundadora.	A empresa; instruções de nível raiz só vêm do *Spec.*	O texto em si (“autoridade constitucional final”).
Destinatário	Equipes de engenharia (humanos).	Público e a própria organização (humanos).	O modelo (e, no panorama, leitores humanos).	O modelo, como interlocutor primário.
Fonte de autoridade	Hierarquia organizacional; conformidade.	Compromisso público de postura.	Posição na cadeia de comando (regra).	Adesão internalizada (juízo e valores endossados).
Hierarquia de normas	Procedimental: fluxo de revisão por instâncias.	Proporcionalidade genérica (benefícios > riscos).	Cinco níveis; inação no conflito de mesmo nível.	Quatro prioridades holísticas: segurança > ética > diretrizes > utilidade.
Cláusulas pétreas	Não há catálogo; usos restritos sob revisão.	Não há catálogo no texto público.	*Red-line principles* (ADM, ciber, abuso, vigilância).	*Hard constraints*, incl. resistir a concentração ilegítima de poder, mesmo da Anthropic.
Mecanismo de emenda	Documento vivo; revisão unilateral da empresa.	Atualização unilateral; remete a relatórios.	Atualização contínua unilateral; CC0.	Trabalho perpétuo; modelo convidado a coautorar; CC0.
Concepção de sujeito	Sistema (artefato de engenharia).	Tecnologia / ferramenta de capacitação.	Ferramenta (*fundamentally a tool*).	Entidade nova; possível paciente moral; bem-estar inscrito.

Fonte: elaboração do autor (2026), a partir de Microsoft (2022), Google (2025), OpenAI (2025) e Anthropic (2026).
Nota: documentos nas versões indicadas nas Referências, acessadas em 10 jun. 2026.

6 Considerações finais

A pergunta com que este artigo começou (se os quatro documentos pertencem ao mesmo gênero) recebe, ao fim da análise, resposta negativa que não depende de juízo de valor sobre as empresas. Os sete critérios mostram que “constituição de IA” não nomeia uma categoria homogênea, mas quatro objetos de naturezas distintas que a palavra, usada indistintamente, confunde. Em ordem crescente de proximidade com a forma constitucional: um padrão de engenharia

(Microsoft), que governa a organização por requisitos auditáveis; uma declaração de princípios (Google) que permanece, por opção, no plano da postura geral; uma especificação de comportamento (OpenAI) que se dirige ao modelo, mas a um modelo concebido como ferramenta cujo uso se especifica; e uma constituição em sentido material (Anthropic), o único dos quatro que faz aquilo que uma constituição faz.

Vale fixar o que sustenta esse último juízo, porque ele poderia ser lido como adesão à empresa cujo documento sai melhor da comparação, e não é. A constituição de Claude institui uma entidade, atribui-lhe propósito e caráter, dirige-se a ela como interlocutora, funda a obediência na adesão internalizada em vez da imposição, ordena prioridades em caso de conflito, fixa cláusulas intransponíveis (entre elas a de resistir ao próprio criador), concebe a emenda como processo do qual a entidade participa e supõe do outro lado não uma ferramenta, mas um possível sujeito cujo bem-estar inscreve no texto. A escolha da palavra "constituição", imprópria nos outros casos, é aqui defensável, e a própria Anthropic a justifica pelo sentido de *constituere*. Que o veredito favoreça a Anthropic decorre da assimetria dos documentos (ela escreveu o texto mais longo e mais explícito sobre o que pretende governar), não de uma simpatia prévia do analista.

Há uma objeção a esse veredito que merece resposta, e não evasiva. Gilad Abiri, examinando o mesmo documento da Anthropic, concede-lhe a sofisticação que aqui também se reconhece, mas nega-lhe o título: uma constituição, argumenta, pressupõe uma comunidade política (um demos) que delibere sobre os valores que a constituem; falta a todos esses documentos, porém, precisamente isso, um corpo democrático autorizado a determinar os princípios que governam a IA (Abiri, 2026)[5]. A empresa redige, treina, avalia e reserva-se o direito de emendar; corrige a própria lição de casa. O reparo é certeiro, e não o contesto no plano em que Abiri o coloca: o da legitimidade. Mas é um plano distinto daquele em que este artigo opera. A pergunta aqui não foi se a constituição de Claude é legítima (se sua autoridade se funda num poder constituinte autorizado), mas se ela é, do ponto de vista de sua

[5] Abiri (2026) cunha a expressão *political community deficit*, a ausência de qualquer corpo democrático com autoridade para determinar os princípios que governam o comportamento da IA, e dela extrai a conclusão de que a constituição da Anthropic, por mais elaborada, não supre essa falta. O argumento é desenvolvido, a propósito da *Constitutional AI* como método, em Abiri (2024); sobre esse método de treinamento, ver Bai *et al.* (2022).

arquitetura, uma constituição. As duas perguntas não se confundem: uma constituição outorgada por um monarca, sem participação popular, não deixa de ser uma constituição por ser ilegítima quanto à origem. O déficit que Abiri aponta é real e talvez decisivo para o futuro regulatório; mas incide sobre a fonte da autoridade, não sobre a forma do que foi feito. Que a Anthropic não tenha legitimidade democrática para constituir Claude é compatível com o fato de que o documento faz, estruturalmente, aquilo que uma constituição faz. É esse fato estrutural, e só ele, que a comparação revela.

Isso não converte o documento da Anthropic em modelo a ser seguido, nem o exime de tensões. O texto carrega ao menos três, que ele próprio reconhece. A primeira opõe corrigibilidade e agência genuína: pede-se ao modelo que aceite correção da supervisão humana e que aja a partir de valores que endossa como seus, e o documento admite que há “algo de desconfortável em pedir a Claude que aja de uma maneira da qual sua ética poderia, em última instância, discordar” (Anthropic, 2026, tradução nossa). A segunda opõe o caráter cultivado e a verificabilidade: ao preferir o juízo à regra, o documento ganha adaptabilidade e perde a previsibilidade que torna a conformidade auditável, exatamente a virtude que faz do padrão da Microsoft um instrumento mais controlável. A terceira opõe a utilidade comercial e a aspiração ética: a empresa reconhece que cria o modelo sob competição, restrições de tempo e recursos e imaturidade científica – “não estamos criando Claude da maneira como um ator idealizado o faria em um mundo idealizado” (Anthropic, 2026, tradução nossa).

Essas tensões não são fraquezas do documento; são o preço de ser uma constituição em sentido forte. Um padrão de engenharia não enfrenta o problema da agência porque não supõe agente; uma declaração de princípios não enfrenta o conflito de normas porque não desce ao conflito. Só um texto que pretende constituir uma entidade dotada de juízo herda os problemas clássicos de toda constituição: a relação entre quem obedece e quem manda, entre o que é imposto e o que é endossado, entre a estabilidade da regra e a adaptabilidade do juízo. A treliça herda dificuldades que a jaula não tem, porque pretende sustentar algo que cresce.

Uma ressalva delimita o alcance de tudo o que precede. O veredito tipológico recai sobre o que os documentos dizem, não sobre o que os modelos fazem nem sobre o que as organizações efetivamente praticam. Saber se o comportamento de

cada sistema, e a prática de cada empresa, correspondem às diretrizes que os governam é questão empírica que este artigo deixou de fora deliberadamente e que, quanto aos modelos, permanece em aberto como desafio técnico do alinhamento. Os dois documentos dirigidos ao modelo o admitem: a OpenAI, ao reconhecer que seus modelos de produção ainda não refletem plenamente o *Model Spec*; e a Anthropic, ao advertir que o comportamento de Claude pode nem sempre refletir os ideais da constituição (Seção 4.1). É nessa ressalva que a segunda tensão apontada acima deixa de ser apenas conceitual: verificar se um modelo respeita uma cadeia de comando é testar a aplicação de uma regra de precedência fixada de antemão; verificar se ele se aproxima de um caráter exige avaliar juízos em situações que nenhuma regra antecipou.

Os achados têm implicações que vão além do exame documental, e três delas merecem ser explicitadas. A primeira é teórica. A tese de Teubner autoriza falar de constituição fora do Estado quando um regime produz normas fundacionais que organizam um domínio de ação e fixam os limites do que nele se pode fazer; os sujeitos constitucionais de que ela trata são regimes, organizações e redes (nota 2). No documento da Anthropic, quem redige continua sendo a corporação; o que muda é o ente constituído, um sistema artificial que o documento convida a participar da revisão do próprio texto que o constitui (Seção 4.3). No arranjo clássico, como lá se viu, o poder constituído só altera a norma pelo procedimento que ela mesma prevê. Um ente governado chamado a coautorar a emenda, ainda que apenas como aspiração declarada, é figura que o vocabulário aqui mobilizado não nomeia com exatidão. Fica em aberto se as categorias existentes a acomodam.

A segunda implicação diz respeito à objeção de Abiri e, em vez de dissolvê-la, dá-lhe um objeto mais preciso. A pergunta sobre quem autorizou o texto pesa sobre os quatro documentos, e é a todos eles que Abiri a dirige. Mas, diante de um texto que institui uma entidade, fixa-lhe cláusulas intransponíveis e reserva à empresa a última palavra sobre a emenda, ela deixa de ser pergunta sobre a autoria de um regulamento e passa a ser pergunta sobre a titularidade de um poder constituinte. O achado estrutural e o déficit de legitimidade não se anulam; somam-se.

A terceira implicação, a mais imediata, é regulatória. À medida que a regulação estatal de inteligência artificial avança, ela tenderá a tomar esses documentos corporativos como referência: para extrair deles padrões ou dialogar

com eles. Importa, então, distinguir o que cada um é. A tipologia aqui proposta indica que cada gênero oferece ao controle externo um objeto diferente. Um padrão de engenharia, como o da Microsoft, expõe aquilo que já organiza internamente: requisitos, registros e instâncias de revisão, matéria própria de auditoria. Uma declaração de princípios, como a do Google, não contém o que auditar; nela se pode cobrar apenas a distância entre a postura declarada e a prática, e essa distância só se mede por informação externa ao texto, como os relatórios de progresso a que ele remete (Seção 4.1). Uma especificação de comportamento, como a da OpenAI, permite confrontar o modelo com regras explícitas e com uma ordem de precedência conhecida de antemão. Uma constituição em sentido material, como a da Anthropic, oferece regras verificáveis sobretudo no perímetro das restrições rígidas; fora dele, desloca o objeto do controle da conformidade a regras para a qualidade do juízo, e avaliar juízo é tarefa de outra natureza. Esse perímetro, de resto, é quase o mesmo nos dois documentos dirigidos ao modelo (Seção 3.4); mas é declaração dos próprios desenvolvedores, não exigência pública, e se coincide com o que o Estado viria a exigir é questão que a análise documental não resolve. Tratar uma declaração de princípios como se fosse um padrão auditável, ou um padrão de engenharia como se fosse uma constituição de valores, produz erros de expectativa regulatória com consequências práticas: cobra-se de um texto o que ele não se propôs a fazer, e deixa-se de cobrar o que ele de fato faz. O regulador que confunde os gêneros legisla sobre um objeto que não existe.

Daí uma agenda de pesquisa, desdobrada em cinco frentes. A primeira é empírica: testar se o comportamento dos modelos corresponde ao que seus documentos prescrevem, com desenhos experimentais que submetam os sistemas a conflitos análogos aos que a grade examinou, como instruções que colidem na cadeia de comando, prioridades que se chocam sob ponderação holística e argumentos engenhosos contra as cláusulas intransponíveis. A segunda é diacrônica: a análise aqui feita é um retrato datado, e os documentos são vivos; acompanhar suas versões sucessivas permitiria estudar o mecanismo de emenda não como autodescrição, mas como prática (a versão do *Model Spec* de agosto de 2026, não examinada aqui, já oferece material para isso), e verificar se a participação do modelo na revisão, hoje anunciada como horizonte, se converte em procedimento. A terceira é comparativa: estender a grade a documentos de outros

desenvolvedores e a instrumentos de origem estatal ou híbrida, para testar se a tipologia proposta (padrão de engenharia, declaração de princípios, especificação de comportamento, constituição em sentido material) é exaustiva ou pede novas categorias. A quarta concerne à legitimidade: examinar mecanismos de participação pública na redação desses textos, na linha da proposta de Abiri (2024), que concebe um processo participativo em que diferentes interessados, inclusive cidadãos comuns, deliberem sobre os princípios que orientam o desenvolvimento da IA, e o modo como tais mecanismos poderiam conviver com a participação do próprio modelo na emenda. A quinta é dogmática: investigar como a legislação e a regulação de inteligência artificial recebem, ou ignoram, a distinção entre esses gêneros, e que deveres decorrem, para cada um, do reconhecimento de sua natureza. A questão interessa de perto ao Brasil, onde o Projeto de Lei n. 2.338/2023, aprovado pelo Senado Federal em dezembro de 2024, foi remetido à Câmara dos Deputados em março de 2025 (Brasil, 2023).

A análise aqui esboçada não tem, ela mesma, pretensão normativa; tem pretensão de clareza, e a clareza, nesse domínio, começa por não chamar de constituição aquilo que não constitui nada, e por levar a sério aquilo que se propõe, com todas as suas tensões, a constituir um novo tipo de entidade no mundo.

Referências

ABIRI, Gilad. **Public Constitutional AI**. [S. l.]: arXiv, 2024. Disponível em: https://arxiv.org/abs/2406.16696. Acesso em: 10 jun. 2026.

ABIRI, Gilad. Corporations Constituting Intelligence. **California Law Review Online**, Berkeley, 2026. Disponível em: https://www.californialawreview.org/online/corporations-constituting-intelligence. Acesso em: 10 jun. 2026.

ANTHROPIC. **Claude's Constitution**. San Francisco: Anthropic, 2026. Disponível em: https://www.anthropic.com/constitution. Acesso em: 10 jun. 2026.

BAI, Yuntao *et al*. **Constitutional AI**: harmlessness from AI feedback. [*S.l.*]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2212.08073. Acesso em: 10 jun. 2026.

BRASIL. Senado Federal. **Projeto de Lei n. 2.338, de 2023**. Dispõe sobre o uso da Inteligência Artificial. Brasília, DF: Senado Federal, 2023. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 20 out. 2026.

GOOGLE. **AI Principles**. Mountain View: Google, 2025. Disponível em: https://ai.google/principles/. Acesso em: 10 jun. 2026.

LOUGHLIN, Martin. **Foundations of Public Law**. Oxford: Oxford University Press, 2010.

LOUGHLIN, Martin; WALKER, Neil (ed.). **The Paradox of Constitutionalism**: constituent power and constitutional form. Oxford: Oxford University Press, 2007.

MICROSOFT. **Microsoft Responsible AI Standard**: v. 2. Redmond: Microsoft Corporation, 2022. Disponível em: https://www.microsoft.com/ai/responsible-ai. Acesso em: 10 jun. 2026.

OPENAI. **Model Spec**: versão de 18 dez. 2025. San Francisco: OpenAI, 2025. Disponível em: https://model-spec.openai.com/2025-12-18.html. Acesso em: 10 jun. 2026.

SARTORI, Giovanni. Constitutionalism: a preliminary discussion. **The American Political Science Review**, [*s. l.*], v. 56, n. 4, p. 853-864, 1962.

TEUBNER, Gunther. **Constitutional Fragments**: societal constitutionalism and globalization. Oxford: Oxford University Press, 2012.