https://www.anthropic.com/legal/commercial-terms

Commercial Terms of Service

Effective June 17, 2025 Previous Version

Welcome to Anthropic! Before accessing our Services, please read these Commercial Terms of Service.

These Commercial Terms of Service ("**Terms**") are an agreement between Anthropic and you or the organization, company, or other entity that you represent ("**Customer**"). "**Anthropic**" means Anthropic Ireland, Limited if Customer resides in the European Economic Area ("**EEA**"), Switzerland or UK, and Anthropic, PBC if Customer resides anywhere else. They govern Customer's use of Anthropic API keys and any other Anthropic offerings that references these Terms, as well as all related Anthropic tools, documentation and services (the "**Services**"). These Terms are effective on the earlier of the date that Customer first electronically consents to a version of these Terms and the date that Customer first accesses the Services ("**Effective Date**").

Please note: You may not enter into these Terms on behalf of an organization, company, or other entity unless you have the legal authority to bind that entity. Services under these Terms are not for consumer use. Our consumer offerings (e.g., Claude.ai) are governed by our Consumer Terms of Service instead.

A. Services

1. **Overview.** Subject to these Terms, Anthropic gives Customer permission to use the Services, including to power products and services Customer makes available to its own customers and end users ("**Users**").
2. **Third Party Features.** Customer may elect (in its sole discretion) to use features, services or other content made available by third parties to Customer through the Services ("**Third Party Features**"). Customer acknowledges and agrees that Third Party Features are not Services and, accordingly, Anthropic is not responsible for them.
3. **Feedback.** If Customer provides (in its sole discretion) Anthropic with feedback regarding the Services, Anthropic may use that feedback at its own risk and without obligation to Customer.

B. Customer Content

As between the parties and to the extent permitted by applicable law, Anthropic agrees that Customer (a) retains all rights to its Inputs, and (b) owns its Outputs. Anthropic disclaims any rights it receives to the Customer Content under these Terms. Subject to Customer's compliance with these Terms, Anthropic hereby assigns to Customer its right, title and interest (if any) in and to Outputs. Anthropic may not train models on Customer Content

from Services. "**Inputs**" means submissions to the Services by Customer or its Users and "**Outputs**" means responses generated by the Services to Inputs (Inputs and Outputs together are "**Customer Content**").

C. Data Privacy

Data submitted through the Services will be processed in accordance with the Anthropic Data Processing Addendum ("**DPA**"), which is incorporated into these Terms by reference.

D. Trust and Safety; Restrictions

1. **Compliance.** Each party will comply with all laws applicable to the provision (for Anthropic) and use (for Customer) of the Services, including any applicable data privacy laws.

2. **Policies and Service Terms.** Customer and its Users may only use the Services in compliance with these Terms, including (a) the Usage Policy ("**Usage Policy**", which was previously referred to as the Acceptable Use Policy), (b) our policy on the countries and regions Anthropic currently supports ("**Supported Regions Policy**") and (c) our Service Specific Terms, each of which is incorporated by reference into these Terms. Customer must cooperate with reasonable requests for information from Anthropic to support compliance with its Usage Policy, including to verify Customer's identity and use of the Services.

3. **Limitations of Outputs; Notice to Users.** It is Customer's responsibility to evaluate whether Outputs are appropriate for Customer's use case, including where human review is appropriate, before using or sharing Outputs. Customer acknowledges, and must notify its Users, that factual assertions in Outputs should not be relied upon without independently checking their accuracy, as they may be false, incomplete, misleading or not reflective of recent events or information. Customer further acknowledges that Outputs may contain content inconsistent with Anthropic's views.

4. **Use Restrictions.** Customer may not and must not attempt to (a) access the Services to build a competing product or service, including to train competing AI models or resell the Services except as expressly approved by Anthropic; (b) reverse engineer or duplicate the Services; or (c) support any third party's attempt at any of the conduct restricted in this sentence.

5. **Service Account.** Customer is responsible for all activity under its account. Customer will promptly notify Anthropic if Customer believes the account it uses to access the Services has been compromised, or is subject to a denial of service or similar malicious attack that may negatively impact the Services.

E. Confidentiality

1. **Confidential Information.** The parties may share information that is identified as confidential, proprietary, or similar, or that a party would reasonably understand to be confidential or proprietary (**"Confidential Information"**). Customer Content is Customer's Confidential Information.

2. **Obligations of Parties.** The receiving party (**"Recipient"**) may only use Confidential Information of the disclosing party (**"Discloser"**) to exercise its rights and perform its obligations under these Terms. Recipient may only share Discloser's Confidential Information to Recipient's employees, agents, and advisors that have a need to know such Confidential Information and who are bound to obligations of confidentiality at least as protective as those provided in these Terms (**"Representatives"**). Recipient will protect Discloser's Confidential Information from unauthorized use, access, or disclosure in the same manner as Recipient protects its own Confidential Information, and with no less than reasonable care. Recipient is responsible for all acts and omissions of its Representatives.

3. **Exclusions.** Confidential Information excludes information that: (a) becomes publicly available through no fault of Recipient; (b) is obtained by Recipient from a third party without a breach of the third party's obligations of confidentiality; or (c) is independently developed by Recipient without use of Confidential Information. Recipient may disclose Discloser's Confidential Information to the extent it is required by law, or court or administrative order, and will, except where expressly prohibited, notify Discloser of the required disclosure promptly and fully cooperate with Discloser's efforts to prevent or narrow the scope of disclosure.

4. **Destruction Request.** Recipient will destroy Discloser's Confidential Information promptly upon request, except where retained to comply with law or copies in Recipient's automated back-up systems, which will remain subject to these obligations of confidentiality while maintained.

F. Intellectual Property

Except as expressly stated in these Terms, these Terms do not grant either party any rights to the other's content or intellectual property, by implication or otherwise.

G. Publicity

Anthropic may use Customer's name and logo to publicly identify Customer as a customer of the Services; provided that Customer may opt-out via this request form. Customer will consider in good faith any request by Anthropic to (1) provide a quote from a Customer executive regarding Customer's motivation for using the Services that Anthropic may use publicly and (2) participate in a public co-marketing activity.

H. Fees

1. **Payment of Fees.** Customer is responsible for fees incurred by its account, at the rates specified on the Model Pricing Page, unless otherwise agreed by the parties. Anthropic may require prepayment for the Services in the form of credits or offer other types of credits, all of which are subject to Anthropic's Supplemental Credits Terms. Anthropic may update the published rates, to be effective the earlier of 30 days after the updates are posted by Anthropic or Customer otherwise receives Notice.

2. **Taxes.** Fees do not include any taxes, duties, or assessments that may be owed by Customer for use of the Services ("**Taxes**"), unless otherwise specified in the applicable invoice. Customer is responsible for remitting any necessary withholding Taxes to the relevant authority on a timely basis and providing Anthropic with evidence of the same upon request. Where law provides for the reduction or elimination of withholding taxes, including via tax treaty, the parties will collaborate in good faith to do so. For clarity, Customer must pay Anthropic the amount ("**Gross-up Payment**") that will ensure that Anthropic receives the same total amount that it would have received if no such withholding or reduction by Customer had been required (taking into account any and all applicable Taxes (including any Taxes imposed on the Gross-up Payment)).

3. **Billing.** Failure to pay Anthropic all amounts owed when due may result in suspension or termination of Customer's access to the Services. Anthropic reserves any other rights of collection it may have.

I. Termination and Suspension

1. **Term.** These Terms start on the Effective Date and continue until terminated (the "**Term**").

2. **Termination.**

    1. Each party may terminate these Terms at any time for convenience with Notice, except Anthropic must provide 30 days prior Notice.

    2. Either party may terminate these Terms for the other party's material breach by providing 30 days prior Notice detailing the nature of the breach unless cured within that time.

    3. Anthropic may terminate these Terms immediately with Notice if Anthropic reasonably believes or determines that Anthropic's provision of the Services to Customer is prohibited by applicable law.

3. **Suspension.**

    1. Anthropic may suspend Customer's access to any portion or all of the Services if: (a) Anthropic reasonably believes or determines that (i) there is a risk to or

attack on any of the Services; (ii) Customer or any User is using the Services in violation of Sections D.1 (Compliance), D.2 (Policies and Service Terms) or D.4 (Use Restrictions); or (iii) Anthropic's provision of the Services to Customer is prohibited by applicable law or would result in a material increase in the cost of providing the Services; or (b) any vendor suspends or terminates Anthropic's use of any third-party services or products required to enable Customer to access the Services (each, a "**Service Suspension**").

2. Anthropic will use reasonable efforts to provide written notice of any Service Suspension to Customer, and resume providing access to the Services, as soon as reasonably possible after the event giving rise to the Service Suspension is cured, where curable. Anthropic will have no liability for any damage, liabilities, losses (including any loss of data or profits), or any other consequences that Customer may incur because of a Service Suspension.

4. **Effect of Termination.** Upon termination, Customer may no longer access the Services. The following provisions will survive termination or expiration of these Terms: (a) Sections E (Confidentiality), G (Publicity), H (Fees), I (Termination and Suspension), J (Disputes), K (Indemnification), L.2 (Disclaimer of Warranties), L.3 (Limits on Liability), and M (Miscellaneous); (b) any provision or condition that must survive to fulfill its essential purpose.

J. Disputes

1. **Disputes.** In the event of a dispute, claim or controversy relating to these Terms ("**Dispute**"), the parties will first attempt in good faith to informally resolve the matter. The party raising the Dispute must notify the other party ("**Dispute Notice**"). The other party will respond to the Dispute Notice in a timely manner. If the parties have not resolved the dispute within 45 days of delivery of the Dispute Notice, either party may seek to resolve the dispute through arbitration as stated in Section J.2 (Arbitration).

2. **Arbitration.** Any Dispute will be determined in English by final, binding arbitration according to the region-specific processes below. Judgment on any award issued through the arbitration process in this Section J.2 (Arbitration) may be entered in any court having jurisdiction. EACH PARTY AGREES THEY ARE WAIVING THE RIGHT TO A TRIAL BY JURY, AND THE RIGHT TO JOIN AND PARTICIPATE IN A CLASS ACTION, TO THE FULLEST EXTENT PERMITTED UNDER THE LAW IN CONNECTION WITH THESE TERMS.

    1. For Customers residing in the EEA, Switzerland or UK, Disputes will be determined by a sole arbitrator in Dublin, Ireland pursuant the UNCITRAL Arbitration Rules as at present in force. The appointing authority shall be the President for the time being of the Law Society of Ireland.

2. For Customers residing anywhere else, Disputes will be determined by a sole arbitrator in San Francisco, CA pursuant to the Comprehensive Arbitration Rules and Procedures of Judicial Arbitration and Mediation Services, Inc.

3. **Equitable Relief.** This Section J (Disputes) does not limit either party from seeking equitable relief.

K. Indemnification

1. **Claims Against Customer.** Anthropic will defend Customer and its personnel, successors, and assigns from and against any Customer Claim (as defined below) and indemnify them for any judgment that a court of competent jurisdiction grants a third party on such Customer Claim or that an arbitrator awards a third party under any Anthropic-approved settlement of such Customer Claim. "**Customer Claim**" means a third-party claim, suit, or proceeding alleging that Customer's paid use of the Services (which includes data Anthropic has used to train a model that is part of the Services) in accordance with these Terms or Outputs generated through such authorized use violates any third-party intellectual property right.

2. **Claims Against Anthropic.** Customer will defend Anthropic and its personnel, successors, and assigns from and against any Anthropic Claim (as defined below) and indemnify them for any judgment that a court of competent jurisdiction grants a third party on such Anthropic Claim or that an arbitrator awards a third party under any Customer-approved settlement of such Anthropic Claim. "**Anthropic Claim**" means any third-party claim, suit, or proceeding related to Customer's or its Users' (a) Inputs or other data provided by Customer, or (b) use of the Services in violation of the Usage Policy, the Service Specific Terms, or Section D.4 (Use Restrictions). Anthropic Claims and Customer Claims are each a "**Claim**", as applicable.

3. **Exclusions.** Neither party's defense or indemnification obligations will apply to the extent the underlying allegation arises from the indemnified party's fraud, willful misconduct, violations of law, or breach of the Agreement. Additionally, Anthropic's defense and indemnification obligations will not apply to the extent the Customer Claim arises from: (a) modifications made by Customer to the Services or Outputs; (b) the combination of the Services or Outputs with technology or content not provided by Anthropic; (c) Inputs or other data provided by Customer; (d) use of the Services or Outputs in a manner that Customer knows or reasonably should know violates or infringes the rights of others; (e) the practice of a patented invention contained in an Output; or (f) an alleged violation of trademark based on use of an Output in trade or commerce.

4. **Process.** The indemnified party must promptly notify the indemnifying party of the relevant Claim, and will reasonably cooperate in the defense. The indemnifying party will retain the right to control the defense of any such Claim, including the selection

of counsel, the strategy and course of any litigation or appeals, and any negotiations or settlement or compromise, except that the indemnified party will have the right, not to be exercised unreasonably, to reject any settlement or compromise that requires that it admit wrongdoing or liability or subjects it to an ongoing affirmative obligation. The indemnifying party's obligations will be excused if either of the following materially prejudices the defense: (a) failure of the indemnified party to provide prompt notice of the Claim; or (b) failure to reasonably cooperate in the defense.

5. **Sole Remedy.** To the extent covered under this Section K (Indemnification), indemnification is each party's sole and exclusive remedy under these Terms for any third-party claims.

L. Warranties and Limits on Liability

1. **Warranties.** Each party represents and warrants that (a) it is authorized to enter into these Terms; and (b) entering into and performing these Terms will not violate any of its corporate rules, if applicable. Customer further represents and warrants that it has all rights and permissions required to submit Inputs to the Services.

2. **Disclaimer of Warranties.** EXCEPT TO THE EXTENT EXPRESSLY PROVIDED FOR IN THESE TERMS, TO THE MAXIMUM EXTENT PERMITTED UNDER LAW (A) THE SERVICES AND OUTPUTS ARE PROVIDED "AS IS" AND "AS AVAILABLE" WITHOUT WARRANTY OF ANY KIND; AND (B) ANTHROPIC MAKES NO WARRANTIES, EXPRESS OR IMPLIED, RELATING TO THIRD-PARTY PRODUCTS OR SERVICES, INCLUDING THIRD-PARTY INTERFACES. ANTHROPIC EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES, INCLUDING WARRANTIES OF MERCHANTABILITY, NON-INFRINGEMENT, AND FITNESS FOR A PARTICULAR PURPOSE, AS WELL AS ANY IMPLIED WARRANTY ARISING FROM STATUTE, COURSE OF DEALING OR PERFORMANCE, OR TRADE USE. ANTHROPIC DOES NOT WARRANT, AND DISCLAIMS THAT, THE SERVICES OR OUTPUTS ARE ACCURATE, COMPLETE OR ERROR-FREE OR THAT THEIR USE WILL BE UNINTERRUPTED. REFERENCES TO A THIRD PARTY IN THE OUTPUTS MAY NOT MEAN THEY ENDORSE OR ARE OTHERWISE WORKING WITH ANTHROPIC.

3. **Limits on Liability.**

    1. Except as stated in Section L.3.b, the liability of each party, and its affiliates and licensors, for any damages arising out of or related to these Terms (i) excludes damages that are consequential, incidental, special, indirect, or exemplary damages, including lost profits, business, contracts, revenue, goodwill, production, anticipated savings, or data, and costs of procurement of substitute goods or services and (ii) is limited to Fees paid by Customer for the Services in the previous 12 months.

2. The limitations of liability in this Section L.3 (Limits on Liability) do not apply to either party's obligations under Section K (Indemnification).

3. THE LIMITATIONS OF LIABILITY IN THIS SECTION L.3 (LIMITS ON LIABILITY) APPLY: (I) TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW; (II) TO LIABILITY IN TORT, INCLUDING FOR NEGLIGENCE; (III) REGARDLESS OF THE FORM OF ACTION, WHETHER IN CONTRACT, TORT, STRICT PRODUCT LIABILITY, OR OTHERWISE; (IV) EVEN IF THE BREACHING PARTY IS ADVISED IN ADVANCE OF THE POSSIBILITY OF THE DAMAGES IN QUESTION AND EVEN IF SUCH DAMAGES WERE FORESEEABLE; AND (E) EVEN IF THE INJURED PARTY'S REMEDIES FAIL OF THEIR ESSENTIAL PURPOSE.

4. The parties agree that they have entered into these Terms in reliance on the terms of this Section L.3 (Limits on Liability) and those terms form an essential basis of the bargain between the parties.

M. Miscellaneous

1. **Notices.** All notices, demands, waivers, and other communications under these Terms (each, a "**Notice**") must be in writing. Except for notices related to demands to arbitrate or where equitable relief is sought, any Notices provided under these Terms may be delivered electronically to the address provided to Anthropic if to Customer; and to notices@anthropic.com if to Anthropic. Notice is effective only: (a) upon receipt by the receiving party, and (b) if the party giving the Notice has complied with all requirements of this Section M.1 (Notices).

2. **Electronic Communications.** Customer agrees to receive electronic communications from Anthropic based on Customer's use of the Services and related to these Terms. Except where prohibited by applicable law, electronic communications may be sent via email, through the Services or Customer's management dashboard, or posted on Anthropic's website. Anthropic may also provide electronic communications via text or SMS about Customer's use of the Services or as Customer otherwise requests from Anthropic. If Customer wishes to stop receiving such messages, Customer may request it from Anthropic or respond to any such texts with "STOP".

3. **Amendment and Modification.** Anthropic may update these Terms at any time, to be effective 30 days after the updates are posted by Anthropic or Customer otherwise receives Notice, except that updates made in response to changes to law or regulation take effect immediately upon posting or Notice. Changes will not apply retroactively. No other amendment to or modification of these Terms is effective unless it is in writing and signed by both parties. Failure to exercise or delay in exercising any rights or remedies arising from these Terms does not and will not be construed as a waiver; and no single or partial exercise of any right or remedy will preclude future exercise of such right or remedy.

4. **Assignment and Delegation.** Neither party may assign its rights or delegate its obligations under these Terms without the other party's prior written consent, except that Anthropic may assign its rights and delegate its obligations to an affiliate or as part of a sale of all or substantially all its business. Any purported assignment or delegation is null and void except as permitted above. No permitted assignment or delegation will relieve the contracting party or assignees of their obligations under these Terms. These Terms will bind and inure to the benefit of the parties and their respective permitted successors and assigns.

5. **Severability.** If a provision of these Terms is invalid, illegal, or unenforceable in any jurisdiction, such invalidity, illegality, or unenforceability will neither affect any other term or provision of these Terms nor invalidate or render unenforceable such term or provision in any other jurisdiction. Upon such determination that any term or other provision is invalid, illegal, or unenforceable, the parties will negotiate in good faith to modify these Terms to reflect the parties' original intent as closely as possible.

6. **Interpretation.** These Terms will be construed mutually, with neither party considered the drafter. Document and section titles are provided for convenience and will not be interpreted. The phrases "for example" or "including" or "or" are not limiting.

7. **Governing Law; Venue.**

    1. These Terms are governed by and construed in accordance with the Governing Laws, without giving effect to any choice of law provision. "**Governing Laws**" means (i) for Customers in the EEA, Switzerland, or UK, the Laws of Ireland; and (ii) for all other Customers, the laws of the State of California.

    2. Any suits, actions, or proceedings related to these Terms that are not required to be resolved via arbitration pursuant to Section J (Disputes) will be instituted exclusively in the Venue, and each party irrevocably submits to their exclusive jurisdiction. "**Venue**" means (i) for Customers in the EEA, Switzerland or UK, the courts of Ireland; and (ii) for all other Customers, federal or state courts located in California.

8. **Export and Sanctions.** Customer may not export or provide access to the Services to persons or entities or into countries or for uses where it is prohibited under U.S. or other applicable international law. Without limiting the foregoing sentence, this restriction applies (a) to countries where export from the US or into such country would be prohibited or illegal without first obtaining the appropriate license, and (b) to persons, entities, or countries covered by U.S. sanctions.

9. **Integration.** These Terms (including the Usage Policy, Supported Regions Policy, Service Specific Terms, DPA, Model Pricing Page and other documents or terms that are incorporated by reference by these Terms) constitute the parties' entire understanding as to the Services' provision and use. These Terms supersede all other understandings or agreements between the parties regarding the Services.
10. **Force Majeure.** Neither party will be liable for failure or delay in performance to the extent caused by circumstances beyond its reasonable control.

https://www.anthropic.com/news/the-long-term-benefit-trust

The Long-Term Benefit Trust

Sep 19, 2023

Today we are sharing more details about our new governance structure called the **Long-Term Benefit Trust (LTBT)**, which we have been developing since the birth of Anthropic. The LTBT is our attempt to fine-tune our corporate governance to address the unique challenges and long-term opportunities we believe transformative AI will present.

The Trust is an independent body of five financially disinterested members with an authority to select and remove a portion of our Board that will grow over time (ultimately, a majority of our Board). Paired with our Public Benefit Corporation status, the LTBT helps to align our corporate governance with our mission of developing and maintaining advanced AI for the long-term benefit of humanity.

Corporate Governance Basics

A corporation is overseen by its board of directors. The board selects and oversees the leadership team (especially the CEO), who in turn hire and manage the employees. The default corporate governance setup makes directors accountable to the stockholders in several ways. For example:

- Directors are elected by, and may be removed by stockholders.
- Directors are legally accountable to stockholders for fulfilling their fiduciary duties.
- Directors are often paid in shares of stock of the corporation, which helps to align their incentives with the financial interests of stockholders.

Importantly, the rights to elect, remove, and sue directors belong exclusively to the stockholders. Some wonder, therefore, whether directors of a corporation are permitted to optimize for stakeholders beyond the corporation's stockholders, such as customers and the general public. This question is the subject of a rich debate, which we won't delve into here. For present purposes, it is enough to observe that all the key mechanisms of accountability in corporate law push directors to prioritize the financial interests of stockholders.

Fine-tuning Anthropic's Corporate Governance

Corporate governance has seen centuries of legal precedent and iteration, and views differ greatly on its effectiveness, strengths, and weaknesses. At Anthropic, our perspective is that the capacity of corporate governance to produce socially beneficial outcomes depends strongly on non-market externalities. Externalities are a type of market failure that occurs when a transaction between two parties imposes costs or benefits on a third party who has not consented to the transaction. Common examples of costs include pollution from factories, systemic financial risk from banks, and national security risks from weapons manufacturers. Examples of positive spillover effects include the societal benefits of education that reach beyond the individuals being educated, or investments in R&D that boost entire sectors beyond the company making the investment. Many parties who contract with a corporation, such as customers, workers, and suppliers, are capable of negotiating or demanding prices and terms that reflect the full costs and benefits of their exchanges. But other parties, such as the general public, don't directly contract with a corporation and therefore do not have a means to charge or pay for the costs and benefits they experience.

The greater the externalities, the less we expect corporate governance defaults to serve the interests of non-contracting parties such as the general public. We believe AI may create unprecedentedly large externalities, ranging from national security risks, to large-scale economic disruption, to fundamental threats to humanity, to enormous benefits to human safety and health. The technology is advancing so rapidly that the laws and social norms that constrain other high-externality corporate activities have yet to catch up with AI; this has led us to invest in fine-tuning Anthropic's governance to meet the challenge ahead of us.

To be clear, for most of the day-to-day decisions Anthropic makes, public benefit is not at odds with commercial success or stockholder returns, and if anything our experience has shown that the two are often strongly synergistic: our ability to do effective safety research depends on building frontier models (the resources for which are greatly aided by commercial success), and our ability to foster a "race to the top" depends on being a viable company in the ecosystem in both a technical sense and a commercial sense. We do not expect the LTBT to intervene in these day-to-day decisions or in our ordinary commercial strategy.

Rather, the need for fine-tuning of the governance structure ultimately derives from the potential for extreme events and the need to handle them with humanity's interests in mind, and we expect the LTBT to primarily concern itself with these long-range issues. For example, the LTBT can ensure that the organizational leadership is incentivized to carefully evaluate future models for catastrophic risks or ensure they have nation-state level security,

rather than prioritizing being the first to market above all other objectives.

Baseline: Public Benefit Corporation

One governance feature we have already shared is that Anthropic is a Delaware Public Benefit Corporation, or PBC. Like most large companies in the United States, Anthropic is incorporated in Delaware, and Delaware corporate law expressly permits the directors of a PBC to balance the financial interests of the stockholders with the public benefit purpose specified in the corporation's certificate of incorporation, and the best interests of those materially affected by the corporation's conduct. The public benefit purpose stated in Anthropic's certificate is the responsible development and maintenance of advanced AI for the long-term benefit of humanity. This gives our board the legal latitude to weigh long- and short-term externalities of decisions–whether to deploy a particular AI system, for example–alongside the financial interests of our stockholders.

The legal latitude afforded by our PBC structure is important in aligning Anthropic's governance with our public benefit mission. But we didn't feel it was enough for the governance challenges we foresee in the development of transformative AI. Although the PBC form makes it legally permissible for directors to balance public interests with the maximization of stockholder value, it does not make the directors of the corporation directly accountable to other stakeholders or align their incentives with the interests of the general public. We set out to design a structure that would supply our directors with the requisite accountability and incentives to appropriately balance the financial interests of our stockholders and our public benefit purpose at key junctures where we expect the consequences of our decisions to reach far beyond Anthropic.

LTBT: Basic Structure and Features

The Anthropic Long-Term Benefit Trust (LTBT, or Trust) is an independent body comprising five Trustees with backgrounds and expertise in AI safety, national security, public policy, and social enterprise. The Trust's arrangements are designed to insulate the Trustees from financial interest in Anthropic and to grant them sufficient independence to balance the interests of the public alongside the interests of Anthropic's stockholders.

At the close of our Series C, we amended our corporate charter to create a new class of stock (Class T) held exclusively by the Trust.1 The Class T stock grants the Trust the authority to elect and remove a number of Anthropic's board members that will phase in according to time- and funding-based milestones; in any event, the Trust will elect a majority of the board within 4 years. At the same time, we created a new director seat that will be elected by the Series C and subsequent investors to ensure that our investors' perspectives will be directly represented on the board into the future.

The Class T stock also includes "protective provisions" that require the Trust to receive

notice of certain actions that could significantly alter the corporation or its business.

The Trust is organized as a "purpose trust" under the common law of Delaware, with a purpose that is the same as that of Anthropic. The Trust must use its powers to ensure that Anthropic responsibly balances the financial interests of stockholders with the interests of those affected by Anthropic's conduct and our public benefit purpose.

A Different Kind of Stockholder

In establishing the Long-Term Benefit Trust we have, in effect, created a different kind of stockholder in Anthropic. Anthropic will continue to be overseen by its board, which we expect will make the decisions of consequence on the path to transformative AI. In navigating these decisions, a majority of the board will ultimately have accountability to the Trust as well as to stockholders, and will thus have incentives to appropriately balance the public benefit with stockholder interests. Moreover, the board will benefit from the insights of Trustees with deep expertise and experience in areas key to Anthropic's public benefit mission. Together we believe the insights and incentives supplied by the Trust will result in better decision making when the stakes are highest.

The gradual "phase-in" of the LTBT will allow us to course-correct an experimental structure and also reflects a hypothesis that, early in a company's history, it can often function best with streamlined governance and not too many stakeholders; whereas as it becomes more mature and has more profound effects on society, externalities tend to manifest themselves progressively more, making checks and balances more critical.

A Corporate Governance Experiment

The Long-Term Benefit Trust is an experiment. Its design is a considered hypothesis, informed by some of the most accomplished corporate governance scholars and practitioners in the nation, who helped our leadership design and "red team" this structure. We're not yet ready to hold this out as an example to emulate; we are empiricists and want to see how it works.

One of the most difficult design challenges was reconciling the imperative for the Trust structure to be resilient to end runs while the stakes are high with the reality of the Trust's experimental nature. It's important to prevent this arrangement from being easily undone, but it is also rare to get something like this right on the first try. We have therefore designed a process for amendment that carefully balances durability with flexibility. We envision that most adjustments will be made by agreement of the Trustees and Anthropic's Board, or the Trustees and the other stockholders. Owing to the Trust's experimental nature, however, we have also designed a series of "failsafe" provisions that allow changes to the Trust and its powers without the consent of the Trustees if sufficiently large supermajorities of the stockholders agree. The required supermajorities increase as the Trust's power phases in, on

the theory that we'll have more experience–and less need for iteration–as time goes on, and the stakes will become higher.

Meet the Initial Trustees
The initial Trustees are:

Jason Matheny: CEO of the RAND Corporation
Kanika Bahl: CEO & President of Evidence Action
Neil Buddy Shah: CEO of the Clinton Health Access Initiative (Chair)
Paul Christiano: Founder of the Alignment Research Center
Zach Robinson: Interim CEO of Effective Ventures US

The Anthropic board chose these initial Trustees after a year-long search and interview process to surface individuals who exhibit thoughtfulness, strong character, and a deep understanding of the risks, benefits, and trajectory of AI and its impacts on society. Trustees serve one-year terms and future Trustees will be elected by a vote of the Trustees. We are honored that this founding group of Trustees chose to accept their places on the Trust, and we believe they will provide invaluable insight and judgment.

[1] An earlier version of the Trust, which was then called the "Long-Term Benefit Committee," was written into our Series A investment documents in 2021, but since the committee was not slated to elect its first director until 2023, we took the intervening time to red-team and improve the legal structure and to carefully consider candidate selection. The current LTBT is the result.

[2] The Trust structure was designed and "red teamed" with immeasurable assistance by John Morley of Yale Law School, David Berger, Amy Simmerman, and other lawyers from Wilson Sonsini, and by Noah Feldman and Seth Berman from Harvard Law School and Ethical Compass Advisors.

Footnotes

In December 2023, Jason Matheny stepped down from the Trust to preempt any potential conflicts of interest that might arise with RAND Corporation's policy-related initiatives. Paul Christiano stepped down in April 2024 to take a new role as the Head of AI Safety at the U.S. AI Safety Institute. Their replacements will be elected by the Trustees in due course.

https://www.anthropic.com/news/anthropics-responsible-scaling-policy

Anthropic's Responsible Scaling Policy

Sep 19, 2023

Today, we're publishing our Responsible Scaling Policy (RSP) – a series of technical and organizational protocols that we're adopting to help us manage the risks of developing increasingly capable AI systems.

As AI models become more capable, we believe that they will create major economic and social value, but will also present increasingly severe risks. Our RSP focuses on catastrophic risks – those where an AI model directly causes large scale devastation. Such risks can come from deliberate misuse of models (for example use by terrorists or state actors to create bioweapons) or from models that cause destruction by acting autonomously in ways contrary to the intent of their designers.

Our RSP defines a framework called AI Safety Levels (ASL) for addressing catastrophic risks, modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials. The basic idea is to require safety, security, and operational standards appropriate to a model's potential for catastrophic risk, with higher ASL levels requiring increasingly strict demonstrations of safety.

A very abbreviated summary of the ASL system is as follows:

- ASL-1 refers to systems which pose no meaningful catastrophic risk, for example a 2018 LLM or an AI system that only plays chess.
- ASL-2 refers to systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't. Current LLMs, including Claude, appear to be ASL-2.
- ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities.
- ASL-4 and higher (ASL-5+) is not yet defined as it is too far from present systems, but will likely involve qualitative escalations in catastrophic misuse potential and autonomy.

The definition, criteria, and safety measures for each ASL level are described in detail in the main document, but at a high level, ASL-2 measures represent our current safety and security standards and overlap significantly with our recent White House commitments. ASL-3 measures include stricter standards that will require intense research and engineering effort to comply with in time, such as unusually strong security requirements and a

commitment not to deploy ASL-3 models if they show any meaningful catastrophic misuse risk under adversarial testing by world-class red-teamers (this is in contrast to merely a commitment to perform red-teaming). Our ASL-4 measures aren't yet written (our commitment is to write them before we reach ASL-3), but may require methods of assurance that are unsolved research problems today, such as using interpretability methods to demonstrate mechanistically that a model is unlikely to engage in certain catastrophic behaviors.

We have designed the ASL system to strike a balance between effectively targeting catastrophic risk and incentivising beneficial applications and safety progress. On the one hand, the ASL system implicitly requires us to temporarily pause training of more powerful models if our AI scaling outstrips our ability to comply with the necessary safety procedures. But it does so in a way that directly incentivizes us to solve the necessary safety issues as a way to unlock further scaling, and allows us to use the most powerful models from the previous ASL level as a tool for developing safety features for the next level.[1] If adopted as a standard across frontier labs, we hope this might create a "race to the top" dynamic where competitive incentives are directly channeled into solving safety problems.

From a business perspective, we want to be clear that our RSP will not alter current uses of Claude or disrupt availability of our products. Rather, it should be seen as analogous to the pre-market testing and safety feature design conducted in the automotive or aviation industry, where the goal is to rigorously demonstrate the safety of a product before it is released onto the market, which ultimately benefits customers.

Anthropic's RSP has been formally approved by its board and changes must be approved by the board following consultations with the Long Term Benefit Trust. In the full document we describe a number of procedural safeguards to ensure the integrity of the evaluation process.

However, we want to emphasize that these commitments are our current best guess, and an early iteration that we will build on. The fast pace and many uncertainties of AI as a field imply that, unlike the relatively stable BSL system, rapid iteration and course correction will almost certainly be necessary.

The full document can be read here. We hope that it provides useful inspiration to policymakers, third party nonprofit organizations, and other companies facing similar deployment decisions.

We thank ARC Evals for their key insights and expertise supporting the development of our RSP commitments, particularly regarding evaluations for autonomous capabilities. We found their expertise in AI risk assessment to be instrumental as we designed our evaluation procedures. We also recognize ARC Evals' leadership in originating and spearheading the

development of their broader ARC Responsible Scaling Policy framework, which inspired our approach.

Footnotes

1. As a general matter, Anthropic has consistently found that working with frontier AI models is an essential ingredient in developing new methods to mitigate the risk of AI.

https://reason.com/2025/06/09/this-ai-company-wants-washington-to-keep-its-competitors-off-the-market/

This AI Company Wants Washington To Keep Its Competitors Off the Market

Anthropic CEO Dario Amodei is petitioning the government to throw roadblocks in his rivals' way.

Jack Nicastro | 6.9.2025 10:44 AM

Share on FacebookShare on XShare on RedditShare by emailPrint friendly versionCopy page URL Add Reason to Google

(Illustration: Eddie Marshall | Inflection AI | Character.AI | Reka AI | Adept | Synthesia | Hertz Foundation | ChatGPT)

Dario Amodei, CEO of the artificial intelligence company Anthropic, published a guest essay in *The New York Times* Thursday arguing against a proposed 10-year moratorium on state AI

regulation. Amodei argues that a patchwork of regulations would be better than no regulation whatsoever.

Skepticism is warranted whenever the head of an incumbent firm calls for more regulation, and this case is no different. If Amodei gets his way, Anthropic would face less competition—to the detriment of AI innovation, AI security, and the consumer.

Amodei's op-ed came in a response to a provision of the so-called One Big Beautiful Bill Act, which would prevent any states, cities, and counties from enforcing any regulation that specifically targets AI models, AI systems, or automated decision systems for 10 years. Senate Republicans have amended the clause from a simple requirement to a condition for receiving federal broadband funds, in order to comply with the Byrd Rule, which in *Politico*'s words "blocks anything but budgetary issues from inclusion in reconciliation."

Amodei begins by describing how, in a recent stress test conducted at his company, a chatbot threatened an experimenter to forward evidence of his adultery to his wife unless he withdrew plans to shut the AI down. The CEO also raises more tangible concerns, such as reports that a version of Google's Gemini model is "approaching a point where it could help people carry out cyberattacks."

Matthew Mittelsteadt, a technology fellow at the Cato Institute, tells *Reason* that the stress test was "very contrived" and that "there are no AI systems where you must prompt it to turn it off." You can just *turn it off*. He also acknowledges that, while there is "a real cybersecurity danger [of] AI being used to spot and exploit cyber-vulnerabilities, it can also be used to spot and *patch*" them.

Outside of cyberspace and in, well, actual space, Amodei sounds the alarm that AI could acquire the ability "to produce biological and other weapons." But there's nothing new about that: Knowledge and reasoning, organic or artificial—ultimately wielded by people in either case—can be used to cause problems as well as to solve them. An AI that can model three-dimensional protein structures to create cures for previously untreatable diseases can also create virulent, lethal pathogens.

Amodei recognizes the double-edged nature of AI and says voluntary model evaluation and publication are insufficient to ensure that benefits outweigh costs. Instead of a 10-year moratorium, Amodei calls on the White House and Congress to work together on a transparency standard for AI companies. In lieu of federal testing standards, Amodei says state laws should pick up the slack without being "overly prescriptive or burdensome." But that caveat is exactly the kind of wishful thinking Amodei indicts proponents of the moratorium for: Not only would 50 state transparency laws be burdensome, says Mittelsteadt, but they could "actually make models *less* legible."

Neil Chilson of the Abundance Institute also inveighed against Amodei's call for state-level regulation, which is much more onerous than Amodei suggests. "The leading state

proposals…include audit requirements, algorithmic assessments, consumer disclosures, and some even have criminal penalties," Chilson tweeted, so "the real debate isn't 'transparency vs. nothing,' but 'transparency-only federal floor vs. intrusive state regimes with audits, liability, and even criminal sanctions.'"

Mittelsteadt thinks national transparency regulation is "absolutely the way to go." But how the U.S. chooses to regulate AI might not have much bearing on Skynet-doomsday scenarios, because, while America leads the way in AI, it's not the only player in the game. "If bad actors abroad create Amodei's theoretical 'kill everyone bot,' no [American] law will matter," says Mittelsteadt. But such a law *can* "stand in the way of good actors using these tools for defense."

A modei is not the only CEO of a leading AI company to call for regulation. In 2023, Sam Altman, co-founder and then-CEO of Open AI, called on lawmakers to consider "intergovernmental oversight mechanisms and standard-setting" of AI. In both cases and in any others that come along, the public should beware of calls for AI regulation that will foreclose market entry, protect incumbent firms' profits from being bid away by competitors, and reduce the incentives to maintain market share the benign way: through innovation and product differentiation.