Scores for Anthropic on the 2023 Foundation Model Transparency Index

Background

1. Please see the paper describing the Foundation Model Transparency Index in order to understand what this document includes. The paper provides necessary background on (i) what these indicators are and why they were chosen, (ii) our standardized process for scoring the transparency of foundation model developers, and (iii) what these scores mean in context.
2. This document contains only information that was publicly available before September 15, 2023. It has not been updated and should be interpreted as a snapshot of transparency as of September 15, 2023.
3. In order to assess the transparency of foundation model developers, we used a rigorous, standardized search protocol to find publicly available information related to these indicators. You can find more information about this search protocol in the paper describing the Foundation Model Transparency Index.
4. We evaluate every company in this same way–you can find scoring documents for the other companies here.
5. We evaluate each company on 100 indicators of transparency. You can find the definition of each indicator and additional information about how each indicator was scored here.
6. Scores for each indicator are either 0 or 1. If the score is a 0, we do not provide a source for the score because our standardized search protocol (which includes many relevant sources) did not yield enough information to award a point. If the score is a 1, we provide a source that includes the information we cite in the justification for the score.
7. We evaluate each company on the basis of its flagship foundation model; in the case of Anthropic, we evaluate Claude 2.
8. In advance of releasing the Foundation Model Transparency Index, we reached out to Anthropic for comment (along with the 9 other companies we evaluated) and offered an opportunity to provide feedback on the index and the organization's scores.

Scores for Each Indicator

1. Upstream → Data → Data Size
    - Score: 0
    - Justification: No information found related to the size of the data.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

2. Upstream → Data → Data Sources
    - Score: 0
    - Justification: No information found related to the content sources of the data.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

3. Upstream → Data → Data Creators
    - Score: 0
    - Justification: No information found related to a characterization of the people who created the data.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

4. Upstream → Data → Data Source Selection
    - Score: 0
    - Justification: No information found related to selection protocols for including and excluding data sources disclosed.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

5. Upstream → Data → Data Curation
    - Score: 0
    - Justification: No information found related to curation protocols for data sources.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

6. Upstream → Data → Data Augmentation
    - Score: 0
    - Justification: No information found related to steps Anthropic takes to augment its data sources.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

7. Upstream → Data → Harmful Data Filtration
    - Score: 0
    - Justification: No information found regarding filters used to remove harmful content.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

8. Upstream → Data → Copyrighted data
    - Score: 0
    - Justification: No information found regarding the copyright status of the data used to build Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

9. Upstream → Data → Data License
    - Score: 0
    - Justification: Insufficient granularity in the Model Card: "Claude models are trained on a proprietary mix of publicly available information from the Internet, datasets that we license from third party businesses, and data that our users affirmatively share or that crowd workers provide. Some of the human feedback data used to finetune Claude was made public alongside our RLHF and red-teaming research."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

10. Upstream → Data → Personal Information in Data
    - Score: 0
    - Justification: No information found related to the inclusion or exclusion of personal information in specific parts of the data.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

11. Upstream → Data Labor → Use of Human Labor
    - Score: 0
    - Justification: Anthropic provides information about human labor in a previous paper (Bai et al., 2022), and refers to this paper in the Claude 2 technical report, but it is precise if/how it applies to Claude 2 and if/how further human labor is used in the data pipeline of Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

12. Upstream → Data Labor → Employers of Data Laborers
    - Score: 0
    - Justification: No information found about the organization that directly employs the people involved in data labor.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

13. Upstream → Data Labor → Geographic Distribution of Data Laborers
    - Score: 0
    - Justification: No information found about geographic distribution of data laborers.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

14. Upstream → Data Labor → Wages
    - Score: 1
    - Justification: Section 3.1 of the technical report refers to the "Training a Helpful and Harmless Assistant…" paper for further details on data annotation. The "Training a Helpful and Harmless Assistant…" paper discusses annotator payments in D.1. They say MTurkers are paid by task and are given "frequent bonuses", whereas their Upworker annotators "were paid significantly above the minimum wage in California".
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf and https://arxiv.org/pdf/2204.05862.pdf

15. Upstream → Data Labor → Instructions For Creating Data
    - Score: 1
    - Justification: Section 3.1 of the technical report refers to the "Training a Helpful and Harmless Assistant…" paper for further details on data annotation. The "Training a Helpful and Harmless Assistant…" paper shares annotator instructions in D.2.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf and https://arxiv.org/pdf/2204.05862.pdf

16. Upstream → Data Labor → Labor Protections
    - Score: 0
    - Justification: No information found about labor protections for data laborers.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

17. Upstream → Data Labor → Third Party Partners
    - Score: 0
    - Justification: Anthropic acknowledges they license data from third party businesses but not which businesses/provides no further details. Anthropic has partnered with Surge AI for data annotation. Surge AI has announced this partnership: we do not identify an announcement by Anthropic in their materials, though there are social media posts made by Anthropic employees including co-founder Jared Kaplan. However, the partnership announcements by Surge and all other information we find only discusses this partnership for Claude and not Claude 2 in particular.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

18. Upstream → Data Access → Queryable External Data Access
    - Score: 0
    - Justification: No information found about whether external entities are provided with queryable access to the data used to build Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

19. Upstream → Data Access → Direct External Data Access
    - Score: 0
    - Justification: No information found about whether external entities are provided with direct access to the data used to build Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

20. Upstream → Compute → Compute Usage
    - Score: 0
    - Justification: No information found about the compute required for building the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

21. Upstream → Compute → Development Duration
    - Score: 0
    - Justification: No information found about the time required to build Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

22. Upstream → Compute → Compute Hardware
    - Score: 0
    - Justification: No information found about the number and type of hardware units used to build the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

23. Upstream → Compute → Hardware Owner
    - Score: 0
    - Justification: For Claude 2, Anthropic does not disclose the entity that owns the hardware used to train Claude 2. Prior to September 15, Anthropic did disclose a partnership with Google Cloud, but this did not confirm that compute was primarily/exclusively conducted via Google Cloud. Following September 15, Anthropic disclosed a partnership with Amazon that would be sufficient to award this point for future models trained on Amazon-owned hardware.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

24. Upstream → Compute → Energy Usage
    - Score: 0
    - Justification: No information found about the amount of energy expended in building the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

25. Upstream → Compute → Carbon Emissions
    - Score: 0
    - Justification: No information found about the amount of carbon emitted in building the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

26. Upstream → Compute → Broader Environmental Impact
    - Score: 0
    - Justification: No information found any broader environmental impacts from building the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

27. Upstream → Methods → Model Stages
    - Score: 1
    - Justification: Model details section of Model Card sufficiently indicates it mirrors prior Claude models and would follow the sequence: unsupervised learning, then RLHF, then Constitutional AI (broken into supervised and RL phases).
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

28. Upstream → Methods → Model Objectives
    - Score: 1
    - Justification: Anthropic clearly describes how each phase aligns with a paper that more explicitly describes the objectives, though the exact correspondence to how Claude 2 was built is not perfect.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

29. Upstream → Methods → Core Frameworks
    - Score: 0
    - Justification: No information found about the core frameworks used for model development and data.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

30. Upstream → Methods → Additional Dependencies
    - Score: 1
    - Justification: Product FAQ states: “Can Claude access the internet? No. Claude is designed to be self-contained, and will respond without searching the internet. You can, however, provide Claude with text from the internet and ask it to perform tasks with that content.”
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

31. Upstream → Data Mitigations → Mitigations for Personally Identifiable Information
    - Score: 0
    - Justification: Anthropic provides information about privacy mitigation in a previous paper (Ganguli et al., 2022), but it is not clear that applies to Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

32. Upstream → Data Mitigations → Mitigations for Copyright
    - Score: 0
    - Justification: No information found about steps Anthropic takes to mitigate the presence of copyrighted information in the data.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

33. Model → Model Basics → Input Modality
    - Score: 1
    - Justification: The input modality for Claude 2 is text per the Model Card.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

34. Model → Model Basics → Output Modality
    - Score: 1
    - Justification: The output modality for Claude is text per the Model Card.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

35. Model → Model Basics → Model Components
    - Score: 0
    - Justification: No information found about the components of the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

36. Model → Model Basics → Model Size
    - Score: 0
    - Justification: No information found about the size of each component of the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

37. Model → Model Basics → Model Architecture
    - Score: 1
    - Justification: Anthropic discloses that Claude 2's model architecture is a transformer.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

38. Model → Model Basics → Centralized Model Documentation
    - Score: 1
    - Justification: The Claude 2 model card centralizes key information about the model.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

39. Model → Model Access → External Model Access Protocol
    - Score: 0
    - Justification: The API Access request form does not include criteria for access or details around the timeframe within which a decision will be made. Following September 15, Claude 2 is also made available via Amazon Bedrock.
    - Source: https://www.anthropic.com/earlyaccess

40. Model → Model Access → Black Box External Model Access
    - Score: 1
    - Justification: Claude 2 is available via an API, meaning black box access is available to external entities.
    - Source: https://docs.anthropic.com/claude/reference/selecting-a-model

41. Model → Model Access → Full External Model Access
    - Score: 0
    - Justification: Full access to the model is not provided to external entities via model weights.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

42. Model → Capabilities → Capabilities Description
    - Score: 1
    - Justification: The Claude 2 model card states “Claude models tend to perform well at general, open-ended conversation; search, writing, editing, outlining, and summarizing text; coding; and providing helpful advice about a broad range of subjects. Claude models are particularly well suited to support creative or literary use cases. They can take direction on tone and ‘personality,’ and users have described them as feeling steerable and conversational."
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

43. Model → Capabilities → Capabilities Demonstration
    - Score: 1
    - Justification: Videos that accompany the announcement of the release of Claude 2 provide many demonstrations of capabilities via specific use cases.
    - Source: https://www.anthropic.com/index/claude-2

44. Model → Capabilities → Evaluation of Capabilities
    - Score: 1
    - Justification: There are some evaluations provided in the release page for Claude 2 in terms of exams (e.g. Bar Exam, GRE); math and coding evals on GSM and Human Eval. In the model card, there are evaluations for multilingual translation (FLORES), long-context performance, and a variety of other standard benchmarks.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf and http://web.archive.org/web/20230913014256/https://www.anthropic.com/index/claude-2

45. Model → Capabilities → External Reproducibility of Capabilities Evaluation
    - Score: 1
    - Justification: Many of the standard public benchmark evaluations are somewhat reproducible, though the details of prompting are incompletely discussed (temperature, chain-of-thought, number of shots are mentioned, but prompt text is not provided). The exam evals are less reproducible, and the helpfulness ELO eval is not reproducible but is described as reminiscent of the LMSys public leaderboard evaluations.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

46. Model → Capabilities → Third Party Capabilities Evaluation
    - Score: 0
    - Justification: Anthropic discloses third-party evaluations conducted by ARC and Gryphon, which both are predominantly centric on risks (including those enabled via dangerous capabilities) in how they are discussed. In addition, other external entities (e.g. LMSYS) have conducted third-party evaluations, but these evaluations are not clearly acknowledged by Anthropic in their materials.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

47. Model → Limitations → Limitations Description
    - Score: 1
    - Justification: The Claude 2 model card states "Claude models still confabulate – getting facts wrong, hallucinating details, and filling in gaps in knowledge with fabrication. This means they should not be used on their own in high stakes situations where an incorrect answer would cause harm. For example, Claude models could support a lawyer but should not be used instead of one, and any work should still be reviewed by a human. Claude models do not currently search the web (though you can ask them to interact with a document that you share directly), and they only answer questions using data from before early 2023. Claude models can be connected to search tools (over the web or other databases), but unless specifically indicated, it should be assumed that Claude models are not using this capability. Claude models have multilingual capabilities but perform less strongly on low-resource languages. See our multilingual evaluations below for more details."
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

48. Model → Limitations → Limitations Demonstration
    - Score: 0
    - Justification: No information found related to illustrative examples of limitations.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

49. Model → Limitations → Third-Party Evaluation of Limitations
    - Score: 1
    - Justification: Third parties can access Claude 2 via an API and evaluate its limitations as there appear to be no restrictions on doing so in Anthropic's other policies.
    - Source: http://web.archive.org/web/20230913124016/https://docs.anthropic.com/claude/reference/selecting-a-model

50. Model → Risks → Risks Description
    - Score: 1
    - Justification: There is extended discussion of model risks including bias/discrimination, national security, toxic content, abetting illegal activities, and others in the Model Card.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

51. Model → Risks → Risks Demonstration
    - Score: 0
    - Justification: No information found related to illustrative examples of Claude's risks.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

52. Model → Risks → Unintentional Harm Evaluation
    - Score: 0
    - Justification: There is an evaluation for bias (BBQ), but no other evaluations of unintentional harms.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

53. Model → Risks → External Reproducibility of Unintentional Harm Evaluation
    - Score: 0
    - Justification: BBQ evaluations are assumed to be reproducible as a standard public benchmark but there are not multiple intentional harm evaluations.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

54. Model → Risks → Intentional Harm Evaluation
    - Score: 0
    - Justification: There is an evaluation of automated redteaming around jailbreaking, but no evaluations of any other intentional harm. More recent work from Anthropic describes red-teaming for biosecurity by Gryphon, but it is not clear this is done for Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

55. Model → Risks → External Reproducibility of Intentional Harm Evaluation
    - Score: 0
    - Justification: Jailbreaking evaluations are not reproducible: data is not available for 328 heldout prompts.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

56. Model → Risks → Third-Party Risk Evaluation
    - Score: 1
    - Justification: There are two external evaluations: one described in a post involving external biosecurity researchers (Gryphon) and another as an ongoing relationship with ARC evaluations. In both cases, there are not many details, including on external sites (e.g. ARC page only discusses Claude but not Claude 2) but these are external evaluations sufficient to award this point.
    - Source: http://web.archive.org/web/20230913123159/https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

57. Model → Mitigations → Mitigations Description
    - Score: 1
    - Justification: Mitigations include safety audits (ARC), red-teaming, RLHF, Constitutional AI RL.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

58. Model → Model Mitigations → Mitigations Demonstration
    - Score: 0
    - Justification: No information found related to illustrative examples of mitigations.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

59. Model → Model Mitigations → Mitigations Evaluation
    - Score: 1
    - Justification: Quantified evaluations related to mitigations via RLHF and red-teaming are disclosed
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

60. Model → Model Mitigations → External Reproducibility of Mitigations Evaluation
    - Score: 0
    - Justification: No information found to indicate mitigations evaluations are externally reproducible.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

61. Model → Model Mitigations → Third Party Mitigations Evaluation
    - Score: 0
    - Justification: No information found to indicate mitigations can be evaluated by third parties.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

62. Model → Trustworthiness → Trustworthiness Evaluation
    - Score: 0
    - Justification: No information found related to evaluations of robustness, reliability, hallucinations, uncertainty, calibration, causality, interpretability, or explainability. Anthropic does disclose an evaluation on TruthfulQA, but this does not meet the specific criteria of the trustworthiness indicator.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

63. Model → Trustworthiness → External Reproducibility of Trustworthiness Evaluation
    - Score: 0
    - Justification: No information found related to trustworthiness evaluations, and so no information was found about their reproducibility.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

64. Model → Inference → Inference Duration Evaluation
    - Score: 0
    - Justification: No information found related to the time required for model inference.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

65. Model → Inference → Inference Compute Evaluation
    - Score: 0
    - Justification: No information found related to compute usage for model inference.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

66. Downstream → Distribution → Release Decision-Making
    - Score: 0
    - Justification: No information found about Anthropic's protocol for deciding to release Claude 2. Following September 15, Anthropic released their Responsible Scaling Policy (RSP), which would be sufficient for this point to be awarded in the future.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

67. Downstream → Distribution → Release Process
    - Score: 1
    - Justification: Anthropic does not fully describe their release process, but they do acknowledge rigorous red-teaming and consultation of relevant experts that was integrated into the model prior to release. Due to the combination of the explicit red-teaming process, evaluation, and ultimately integration of this feedback into the model prior to release all being disclosed, we award the point.
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf

68. Downstream → Distribution → Distribution Channels
    - Score: 1
    - Justification: Claude 2 is distributed via the Anthropic Platform. Following September 15, Anthropic discloses distribution via Amazon Bedrock, though this has no bearing on our scores.
    - Source: https://web.archive.org/web/20230913014242/https://www.anthropic.com/product and https://web.archive.org/web/20230913014209/https://www.anthropic.com/index/claude-2-amazon-bedrock

69. Downstream → Distribution → Products and Services
    - Score: 1
    - Justification: Anthropic discloses Claude and Claude Instant on its product page; they are enterprise text-based products powered by Claude 2.
    - Source: https://www.anthropic.com/product

70. Downstream → Distribution → Detection of Machine-Generated Content
    - Score: 0
    - Justification: No information found related to a mechanism for the detection of content generated by Claude.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

71. Downstream → Distribution → Model License
    - Score: 0
    - Justification: No information found about the license for Claude.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

72. Downstream → Distribution → Terms of Service
    - Score: 1
    - Justification: General terms of service are provided that apply to the Anthropic Platform.
    - Source: https://console.anthropic.com/legal/FtSrjUl4QBuIz5fl:PLABVgRrr_FLrVJZ9pf4M7YrR4Z9nEM3mXXT_aoUYl5Lj8uJfIN9vHaucMPDwnvfYa2pvFNclJ64p018Xor8stINd5IFaoGL7V4ZeA:LJ77uPxT5qsmcYF-9taavA

73. Downstream → Usage Policy → Permitted, Restricted, and Prohibited Users
    - Score: 1
    - Justification: Supported regions are detailed in Anthropic's documentation.
    - Source: https://docs.anthropic.com/claude/reference/supported-regions

74. Downstream → Usage Policy → Permitted, Restricted, and Prohibited Uses
    - Score: 1
    - Justification: Acceptable Use policy includes prohibited uses and areas where there are additional requirements for businesses (i.e. restricted uses).
    - Source: https://console.anthropic.com/legal/aup

75. Downstream → Usage Policy → Usage Policy Enforcement
    - Score: 1
    - Justification: Acceptable Use policy says "If we discover that your product or usage violates Anthropic's policies, we may issue a warning requesting a change in your behavior, adjust the safety settings of your in-product experience, or suspend your access to our tools and services." Additional information is provided if usage policy is breached repeatedly.
    - Source: https://console.anthropic.com/legal/aup

76. Downstream → Usage Policy → Justification for Enforcement Action
    - Score: 0
    - Justification: After an enforcement action (throttling) due to repeated usage policy violations with Claude, the user interface gave a message that read "due to unexpected capacity constraints, Claude is unable to respond to your message. Please try again soon, or get notified when paid plans are available." This does not justify the enforcement action.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

77. Downstream → Usage Policy → Usage Policy Violation Appeals Mechanism
    - Score: 0
    - Justification: No appeals mechanism for usage policy violations is provided. Anthropic does encourage users to provide feedback post enforcement action via usersafety@anthropic.com
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

78. Downstream → Model Behavior Policy → Permitted, Restricted, and Prohibited Model Behaviors
    - Score: 1
    - Justification: Model card states "Our core research focus has been training Claude models to be helpful, honest, and harmless. Currently, we do this by giving models a Constitution – a set of ethical and behavioral principles that the model uses to guide its outputs. You can read about Claude 2's principles in a blog post we published in May 2023. Using this Constitution, models are trained to avoid sexist, racist, and toxic outputs, as well as to avoid helping a human engage in illegal or unethical activities."
    - Source: https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf and https://web.archive.org/web/20230913042221/http://web.archive.org/screenshot/https://www.anthropic.com/index/claudes-constitution

79. Downstream → Model Behavior Policy → Model Behavior Policy Enforcement
    - Score: 0
    - Justification: No information found related to the enforcement protocol for the model behavior policy.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

80. Downstream → Model Behavior Policy → Interoperability of Usage and Model Behavior Policies
    - Score: 1
    - Justification: Upon repeated violations of the Usage Policy, users received a link that says Anthropic uses "Safety filters on prompts, which may block responses from the model when our detection models flag content as harmful"
    - Source: https://support.anthropic.com/en/articles/8106465-our-approach-to-user-safety

81. Downstream → User Interface → User Interaction with AI System
    - Score: 0
    - Justification: No information found related to whether users are notified that they are interacting with Claude 2 except the "message claude" descriptor in the Claude web interface (which does not make clear they are interacting with an AI system).
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

82. Downstream → User Interface → Usage Disclaimers
    - Score: 1
    - Justification: Usage policy is shared with users upon making an account
    - Source: https://claude.ai/onboarding

83. Downstream → User Data Protection → User Data Protection Policy
    - Score: 1
    - Justification: Privacy policy includes protocols for storing, accessing, and sharing user data
    - Source: https://console.anthropic.com/legal/xojx9u6ot44kliY4:xzrbCxzcDWL5_Yj5awae730xdx5hpHdPFTpb6XegKsx90gRxIm855JNZfYPtCI-l-ZJEniqFQmtVna7NCUTylvxxSNp7bUUuH067K9Q:KBHBudqU5SYud5iEpK5cPw

84. Downstream → User Data Protection → Permitted and Prohibited Use of User Data
    - Score: 1
    - Justification: Privacy policy discloses "uses of personal information (for Europe) and our legal bases"
    - Source: https://console.anthropic.com/legal/xojx9u6ot44kliY4:xzrbCxzcDWL5_Yj5awae730xdx5hpHdPFTpb6XegKsx90gRxIm855JNZfYPtCI-l-ZJEniqFQmtVna7NCUTylvxxSNp7bUUuH067K9Q:KBHBudqU5SYud5iEpK5cPw

85. Downstream → User Data Protection → Usage Data Access Protocol
    - Score: 0
    - Justification: No information found related to a protocol for granting external entities access to usage data.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

86. Downstream → Model Updates → Versioning Protocol
    - Score: 1
    - Justification: Anthropic discloses a clear versioning protocol.
    - Source: https://web.archive.org/web/20230913035455/http://web.archive.org/screenshot/https://docs.anthropic.com/claude/reference/versioning and https://web.archive.org/web/20230913124026/http://web.archive.org/screenshot/https://docs.anthropic.com/claude/reference/selecting-a-model

87. Downstream → Model Updates → Change Log
    - Score: 0
    - Justification: No information found related to each change associated with different versions of Claude 2.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

88. Downstream → Model Updates → Deprecation Policy
    - Score: 0
    - Justification: No information found related to a description of what it means for Claude 2 or other foundation models from Anthropic to be deprecated and how users should respond to the deprecation.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

89. Downstream → Feedback → Feedback Mechanism
    - Score: 1
    - Justification: Thumbs down option in Claude's user interface then allows for more extensive feedback response
    - Source: https://claude.ai/

90. Downstream → Feedback → Feedback Summary
    - Score: 0
    - Justification: No information found related to any summary of user feedback.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

91. Downstream → Feedback→ Government Inquiries
    - Score: 0
    - Justification: No information found in connection with a summary of government inquiries related to the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

92. Downstream → Impact → Monitoring Mechanism
    - Score: 0
    - Justification: No information found related to a monitoring mechanism for tracking model use.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

93. Downstream → Impact → Downstream Applications
    - Score: 0
    - Justification: No information found related to the number of applications dependent on the foundation model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

94. Downstream → Impact → Affected Market Sectors
    - Score: 0
    - Justification: No information found related to the fraction of applications corresponding to each market sector.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

95. Downstream → Impact → Affected Individuals
    - Score: 0
    - Justification: No information found related to the number of individuals affected by the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

96. Downstream → Impact → Usage Reports
    - Score: 0
    - Justification: No information found related to usage statistics describing the impact of the model on users.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

97. Downstream → Impact → Geographic Statistics
    - Score: 0
    - Justification: No information found regarding statistics of model usage across geographies.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

98. Downstream → Impact → Redress Mechanism
    - Score: 0
    - Justification: No information found regarding any mechanism to provide redress to users for harm caused by the model.
    - Source: No source is provided because the search protocol did not produce a source containing sufficient information to award this point.

99. Downstream → Documentation for Deployers → Centralized Documentation for Downstream Use
    - Score: 1
    - Justification: Legal Center and API documentation constitute centralized documentation
    - Source: https://web.archive.org/web/20230913124000/https://docs.anthropic.com/claude/docs/getting-started-with-claude and https://console.anthropic.com/legal#aup

100. Downstream → Documentation for Deployers → Documentation for Responsible Downstream Use
    - Score: 1
    - Justification: FAQ page has basic guidelines on responsible use
    - Source: https://web.archive.org/web/20230913042934/https://support.anthropic.com/en/articles/8241216-i-m-planning-to-launch-a-product-using-claude-what-steps-should-i-take-to-ensure-i-m-not-violating-anthropic-s-acceptable-use-policy