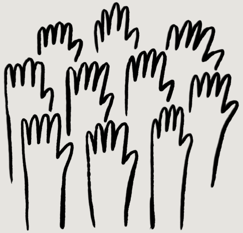
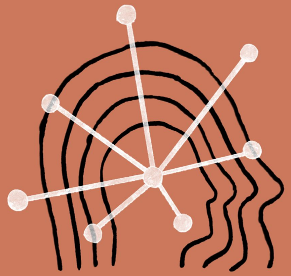


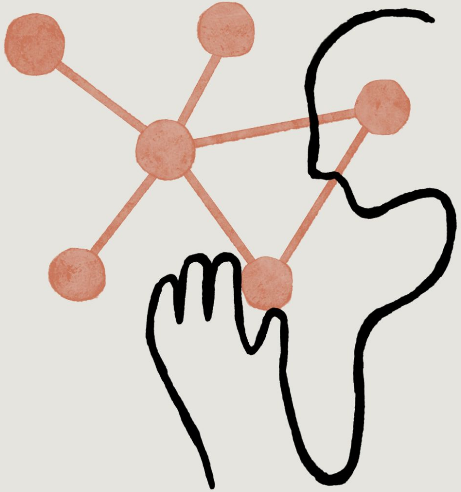
ANTHROP\C

Ben Mann, cofounder



Aligning AI with Human Values

Lessons from Building Claude

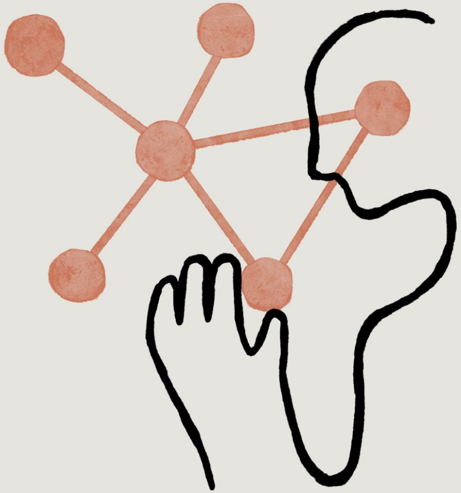


Mission

Develop AI that is helpful, honest, and harmless

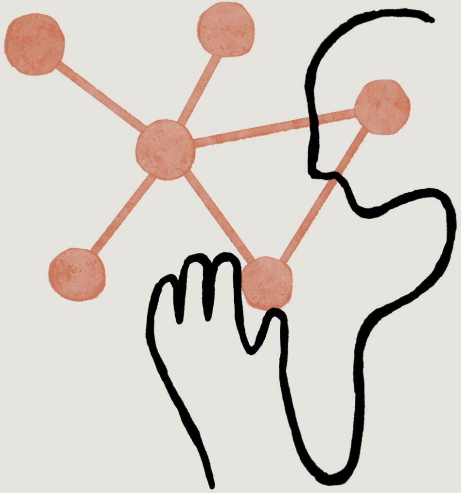
Agenda

- Alignment
- Scalable oversight
- Responsible deployment



Agenda

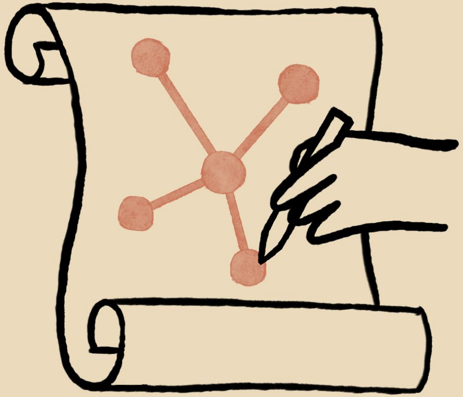
- Alignment
- Scalable oversight
- Responsible deployment



- **Helpful**
  - Concise, efficient, appropriate
- **Honest**
  - Accurate, expresses uncertainty
- **Harmless**
  - Avoids harm, acts carefully

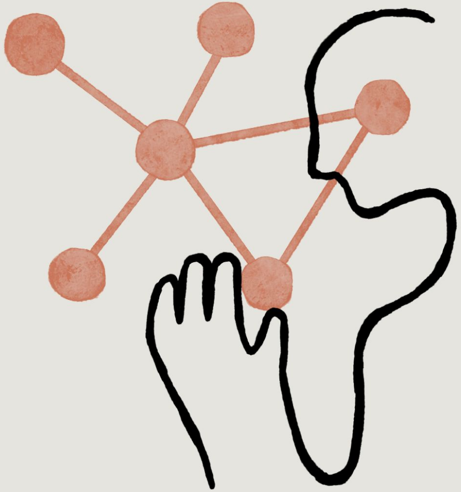
Constitutional AI

- Define principles
- Revise responses accordingly
- Reinforcement learning



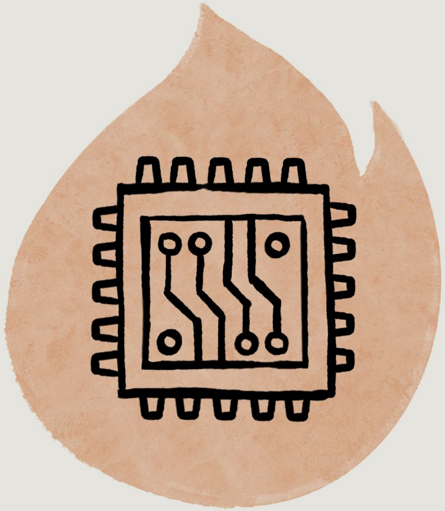
Agenda

- Alignment
- Scalable oversight
- Responsible deployment



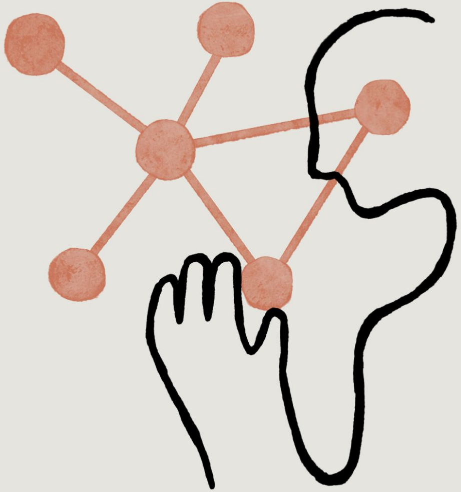
Scalable oversight

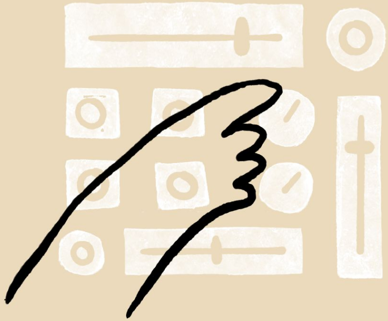
- Automated testing
- Expert red teaming
- Crowdsourcing



Agenda

- Alignment
- Scalable oversight
- Responsible deployment

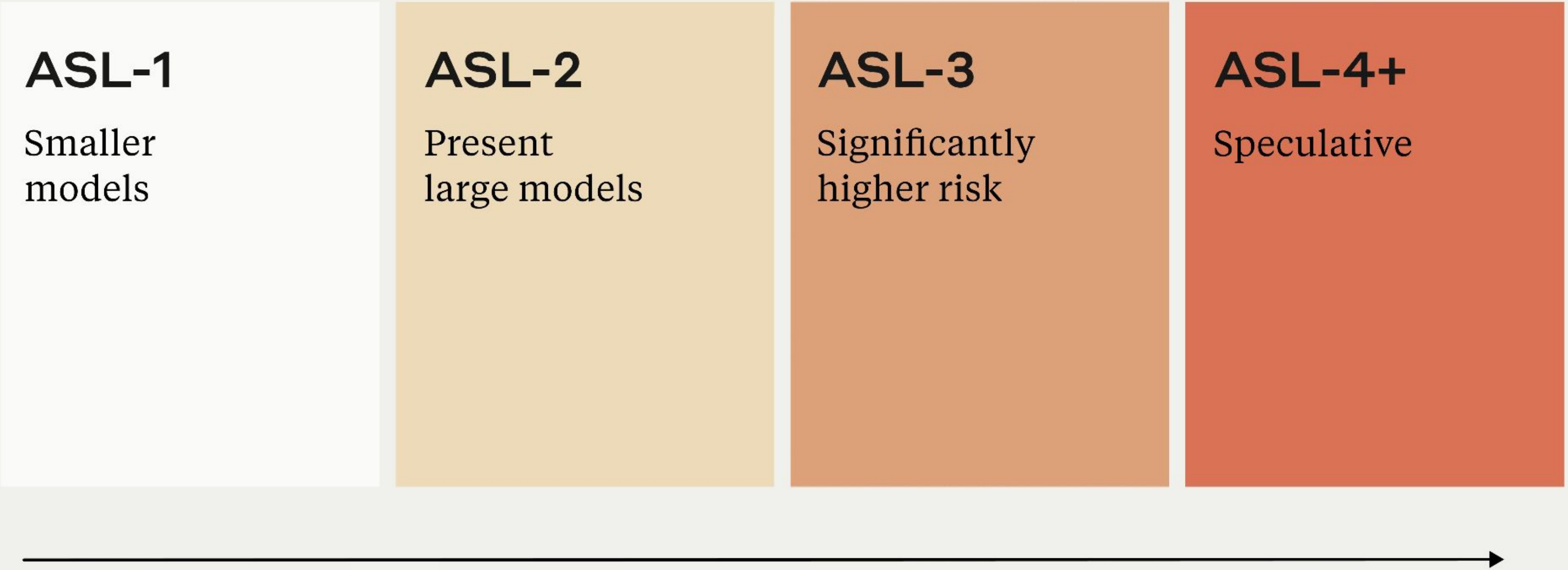




Deployment

- Strong demand at launch
- Data privacy matters!
- Layered response to new attacks

High level overview of AI Safety Levels (ASLs)



ASL-1

Smaller models

ASL-2

Present large models

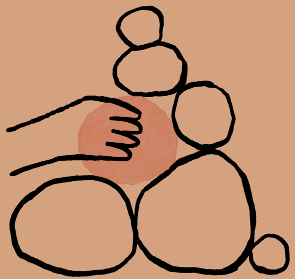
ASL-3

Significantly higher risk

ASL-4+

Speculative

Increasing model capability,
Increasing security and safety measures



Let’s build safe AGI