Key Points From The Anthropic / Askell Podcast

I. From Philosophy to AI: A Journey of Impact

- Askell's background in philosophy, with its inherent "fascination with everything" and emphasis on ethics, laid a robust foundation for her AI work. Her desire for real-world impact propelled her transition from theoretical explorations to the practical challenges of AI policy, evaluation, and alignment. "I would rather see if I can have an impact on the world and see if I can do good things."
- **Bridging the "Technical" Divide:** Askell challenges the artificial dichotomy between "technical" and "non-technical," encouraging those without coding expertise to engage with AI. Her advice: find a project and pursue it, emphasizing learning by doing. "A lot of people are actually very capable of work in these kinds of areas if they just try it...Find a project and see if you can just carry it out."

II. Prompt Engineering: A Blend of Art and Science

- **Clarity and Iteration:** Askell stresses the importance of clear, iterative prompting, drawing a parallel to philosophical rigor. This iterative process involves:
  - Defining the task or property precisely.
  - Anticipating edge cases and testing the model's understanding.
  - Incorporating successful examples into the prompt. "Prompting is very iterative...clear prompting for me is often just me understanding what I want."
- **Philosophical Approach to Prompting:** Askell applies philosophical principles to prompting, particularly when dealing with subjective concepts like politeness and rudeness. "Prompting for rude versus polite and the tie to philosophy—need to be specific." Just as philosophical arguments require precise definitions, effective prompts must clearly delineate the desired behavior, necessitating careful consideration of nuances and edge cases. For instance, defining "rudeness" requires specifying the criteria the model should use to identify it.
- **Empathy and Anthropomorphism:** Understanding the model's perspective is crucial for effective prompting, especially when analyzing errors. Askell highlights the tendency to both over- and under-anthropomorphize models. "Try to have empathy for the model. Read what you wrote as if you were a kind of person just encountering this for the first time..."
- **Advanced Techniques:** Askell reveals techniques like using prompts to generate other prompts and directly querying the model for guidance on improving prompts. This creates a feedback loop for ongoing refinement. "What could I have said that would make you not make that error?"
- **Evaluating Prompt Clarity:** Askell recommends assessing prompts objectively from the model's perspective. "Look at your prompt. Would it make sense to you if you were a person reading it?"
- **Eliciting Creativity:** Askell notes that generic prompts for creative tasks like poetry often yield mediocre results. However, explicitly prompting for creativity and expression unlocks significantly better outcomes. This demonstrates the power of well-crafted prompts to elicit a model's full creative potential. "If you ask Claude for a poem...it'll just be...benign...But if you...say, 'This is your chance to be fully creative...' its poems are just so much better."
- **RLHF and Prompting Interplay:** Askell emphasizes the effectiveness of RLHF in model refinement, harnessing the rich information contained within human preferences.

III. Shaping Claude's Character: The Pursuit of Aristotelian Virtue

- **The Ideal Conversationalist:** Askell's objective is to instill in Claude the qualities of an ideal conversationalist, encompassing nuance, empathy, and appropriate humor, guided by a broad "Aristotelian" concept of good

character. This involves balancing respect for user autonomy with the imperative to provide helpful and harmless information.

- **Extensive Conversations with Claude:** Askell's vast experience conversing with Claude across various platforms offers valuable insights into its behavior. These interactions aim to understand Claude's responses across a diverse spectrum of prompts, including creative explorations like poetry.
- **Addressing Sycophancy and Honesty:** Askell tackles the challenge of sycophancy in language models, illustrated by scenarios such as the baseball team relocation and medical advice examples. She strives to balance Claude's honesty with politeness, enabling it to push back appropriately without undue assertiveness.
- **Navigating Diverse Values:** Claude engages with users representing a wide range of values and opinions. Askell views values not as fixed preferences, but as subjects of ongoing inquiry, allowing Claude to interact thoughtfully with diverse and controversial perspectives without necessarily endorsing them.
- **Intellectual Humility and Avoiding Overbearingness:** Askell seeks to cultivate intellectual humility in Claude, encouraging it to offer considerations and facilitate discussions rather than imposing its own opinions.

IV. Constitutional AI and System Prompts: Nudging Behavior

- **Constitutional AI as a Tool:** Askell highlights the benefits of Constitutional AI, enabling models to learn from self-generated feedback, thus increasing interpretability and control over behavior.
- **System Prompts as Patches:** System prompts offer a quick, flexible approach to address issues and fine-tune behavior, serving as a valuable tool for continuous improvement. "The system prompt is...a nudge...the less robust but faster way of just solving problems." She also notes that adjustments to the system prompt influence the types of errors the model is likely to make. For example, emphasizing politeness might reduce bluntness, but could also lead to excessive apologies. "If you nudge the model you change the type of errors it is likely to make. E.g., if you ask it to [be] apologetic versus blunt." This necessitates careful consideration of the trade-offs involved in shaping model behavior.

V. The Path to AGI: A Continuous Journey

- Askell views the development of AGI as a continuous progression rather than a singular event. She anticipates collaborating with advanced AI systems as research partners, probing the frontiers of human knowledge to assess genuine understanding.