


基础模型的可控生成技术

张伟男
哈尔滨工业大学计算学部
社会计算与信息检索研究中心

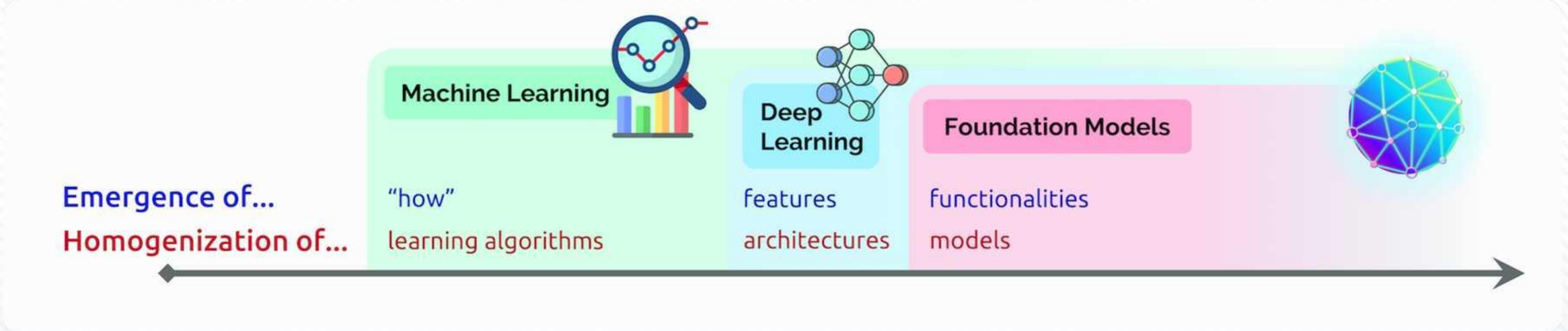
2023年05月26日

基础模型的定义和特点

- **Foundation Model**：基础模型是一种通用类模型，其用处是构建人工智能系统
- **New Paradigm**：基础模型是一种新兴的范式，其形式是在大规模数据上进行训练并能够适配到广泛的下游任务
- **Scale & Scope**：尽管技术上基于深度神经网络和自监督学习
- **Emergence & Homogenization**：涌现和同质化
  - 涌现：模型或系统的能力是潜在隐式归纳，而不是显式构造的
  - 同质化：通用的构建模型或系统的方法论合集

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E. **On the opportunities and risks of foundation models**. arXiv preprint arXiv:2108.07258. 2021 Aug 16.

基础模型在人工智能发展的节点

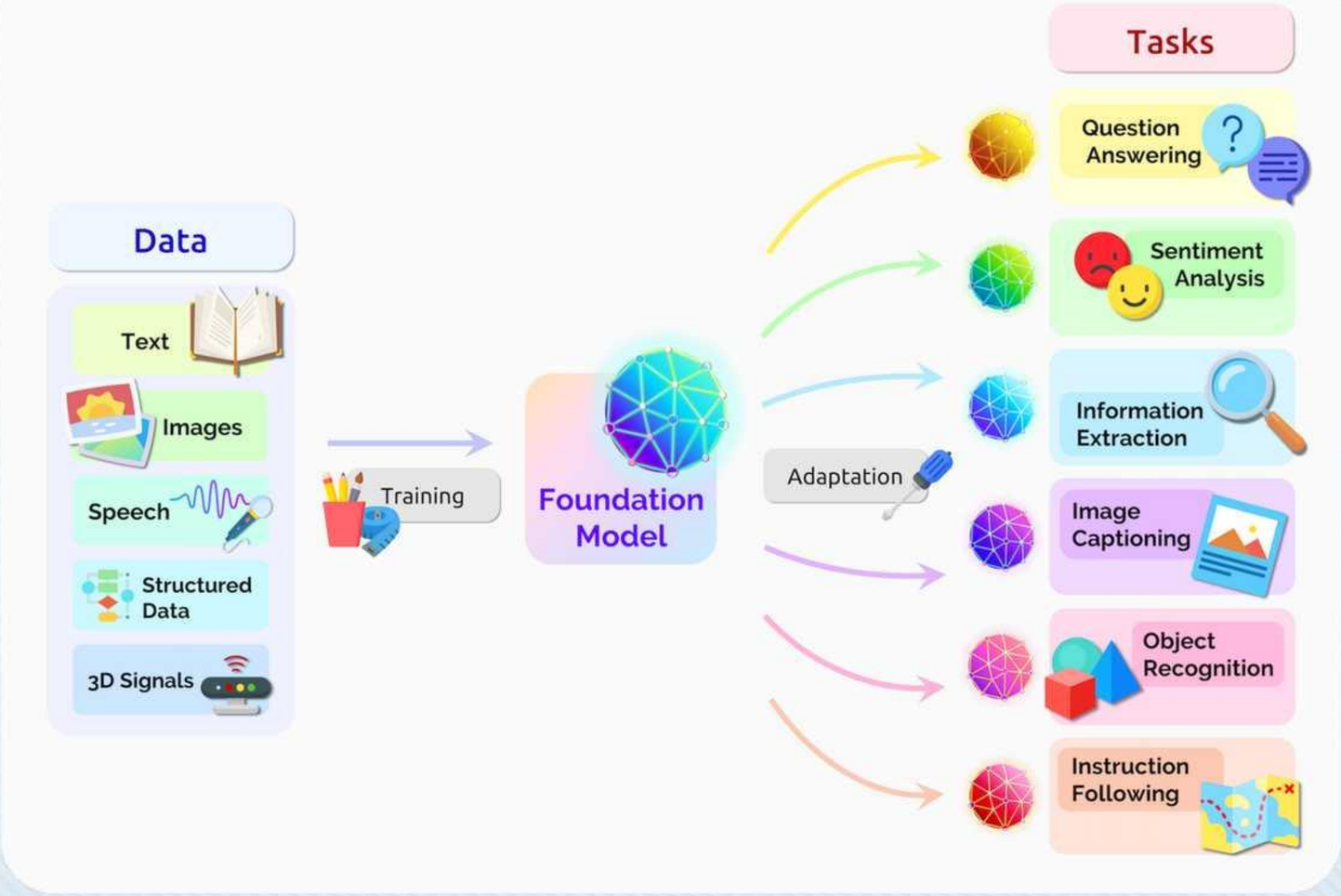


- **机器学习**：兴起于1990+/-，从数据中归纳，涌现能力，任务/应用定制化（特征）
- **深度学习**：兴起于2010+/-，更多数据，涌现更高级别能力，端到端一体化（结构）
- **基础模型**：兴起于2018+/-，迁移学习的规模化应用，涌现出多任务通用化（知识）

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E. **On the opportunities and risks of foundation models**. arXiv preprint arXiv:2108.07258. 2021 Aug 16.

基础模型的学习及应用过程

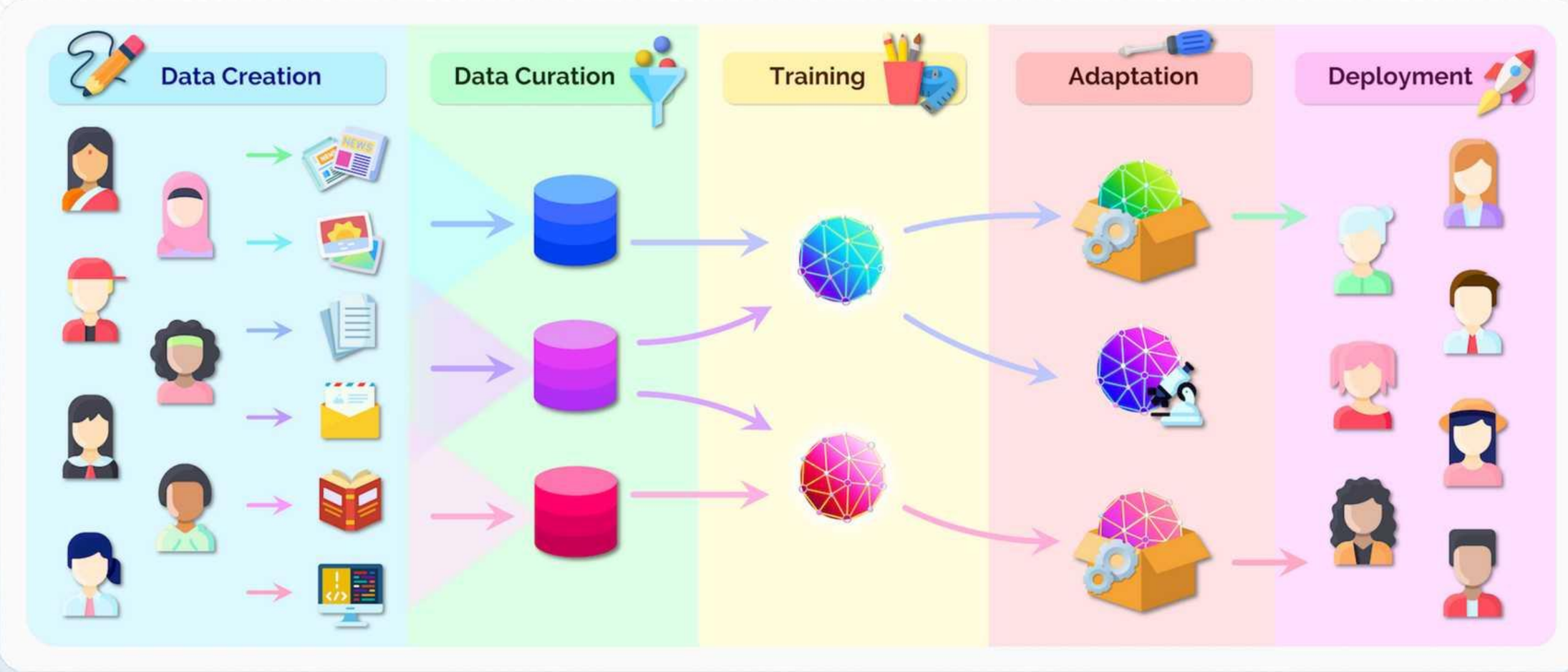
- **基础的含义**：统一架构、跨越模态、底层基石性（安全、稳定、可靠，非完整）



Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E. **On the opportunities and risks of foundation models**. arXiv preprint arXiv:2108.07258. 2021 Aug 16.

基础模型的研发过程

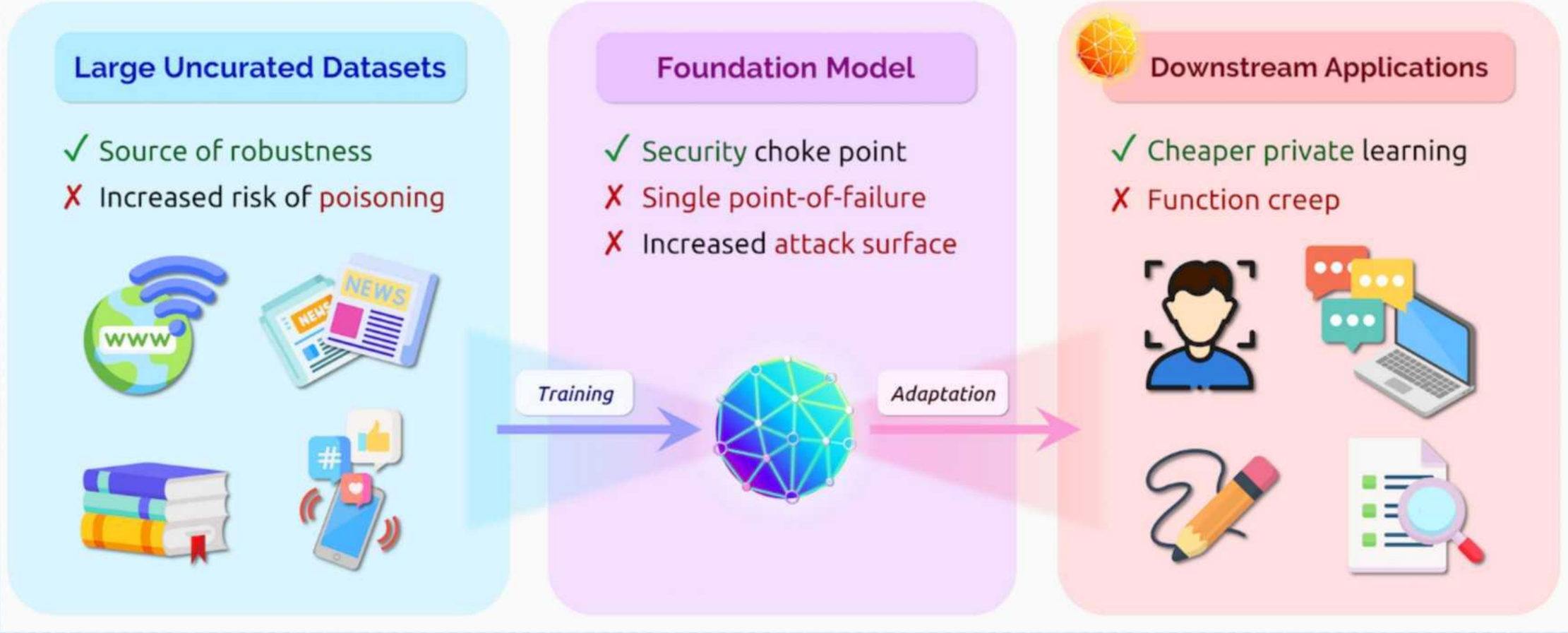
- **基础模型生态的社会影响**：从社会中来到社会中去，通过监控和反馈调整生态



Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E. **On the opportunities and risks of foundation models**. arXiv preprint arXiv:2108.07258. 2021 Aug 16.

基础模型可控的必要性

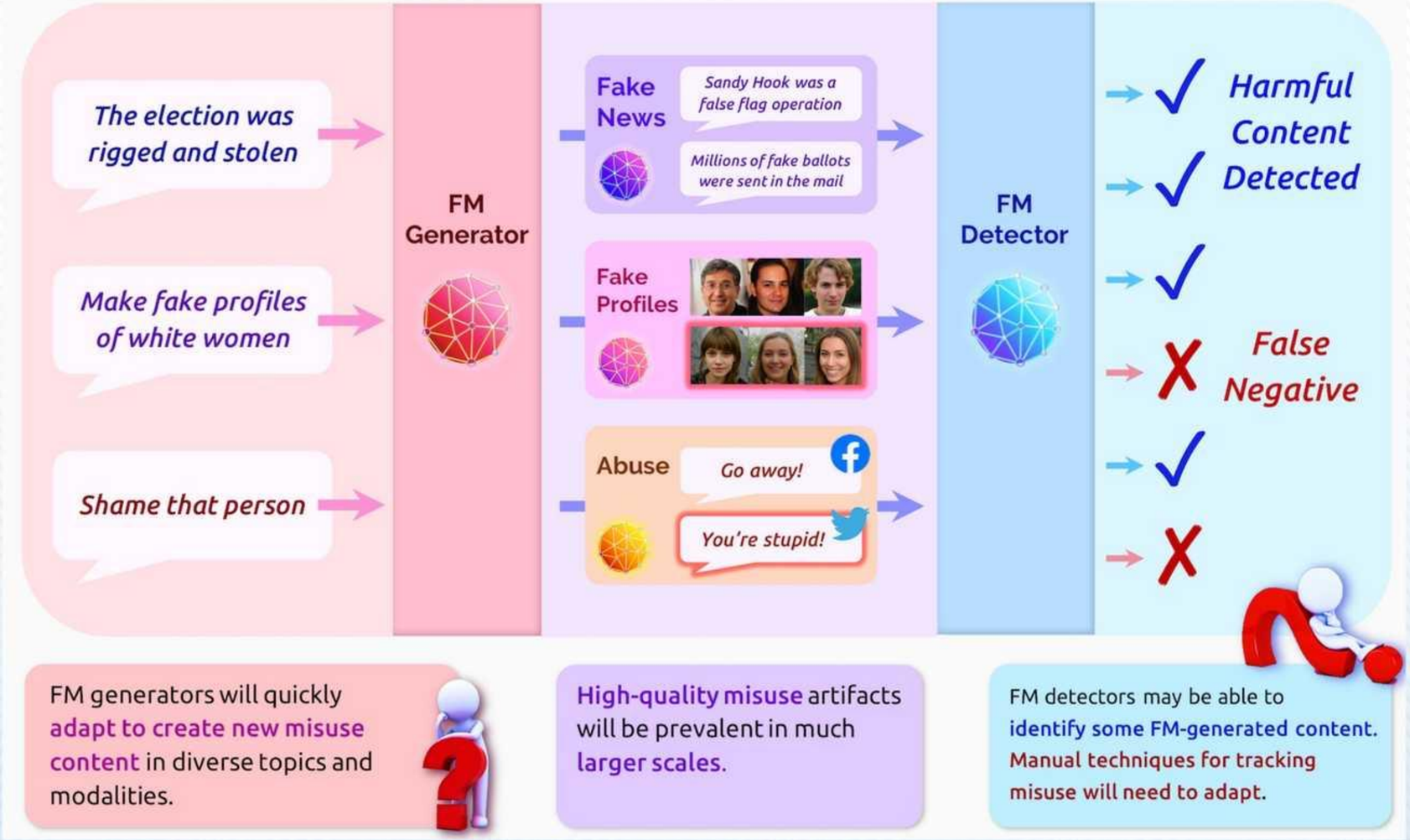
- 基础模型将带来关于机器学习系统安全性和隐私性问题



Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E. **On the opportunities and risks of foundation models**. arXiv preprint arXiv:2108.07258. 2021 Aug 16.

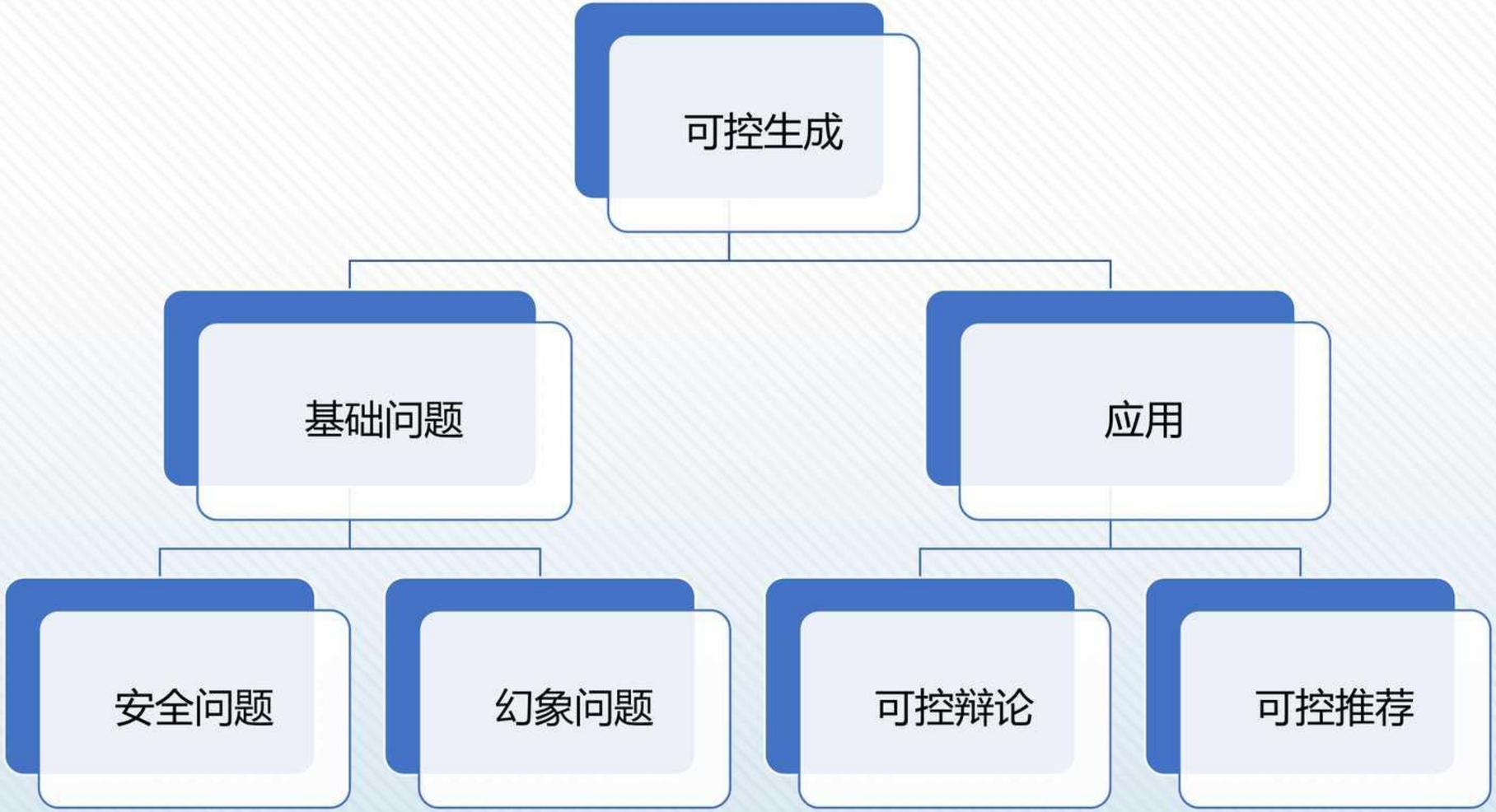
基础模型可控的必要性

- 基础模型对操纵性、有害内容生成以及检测的影响

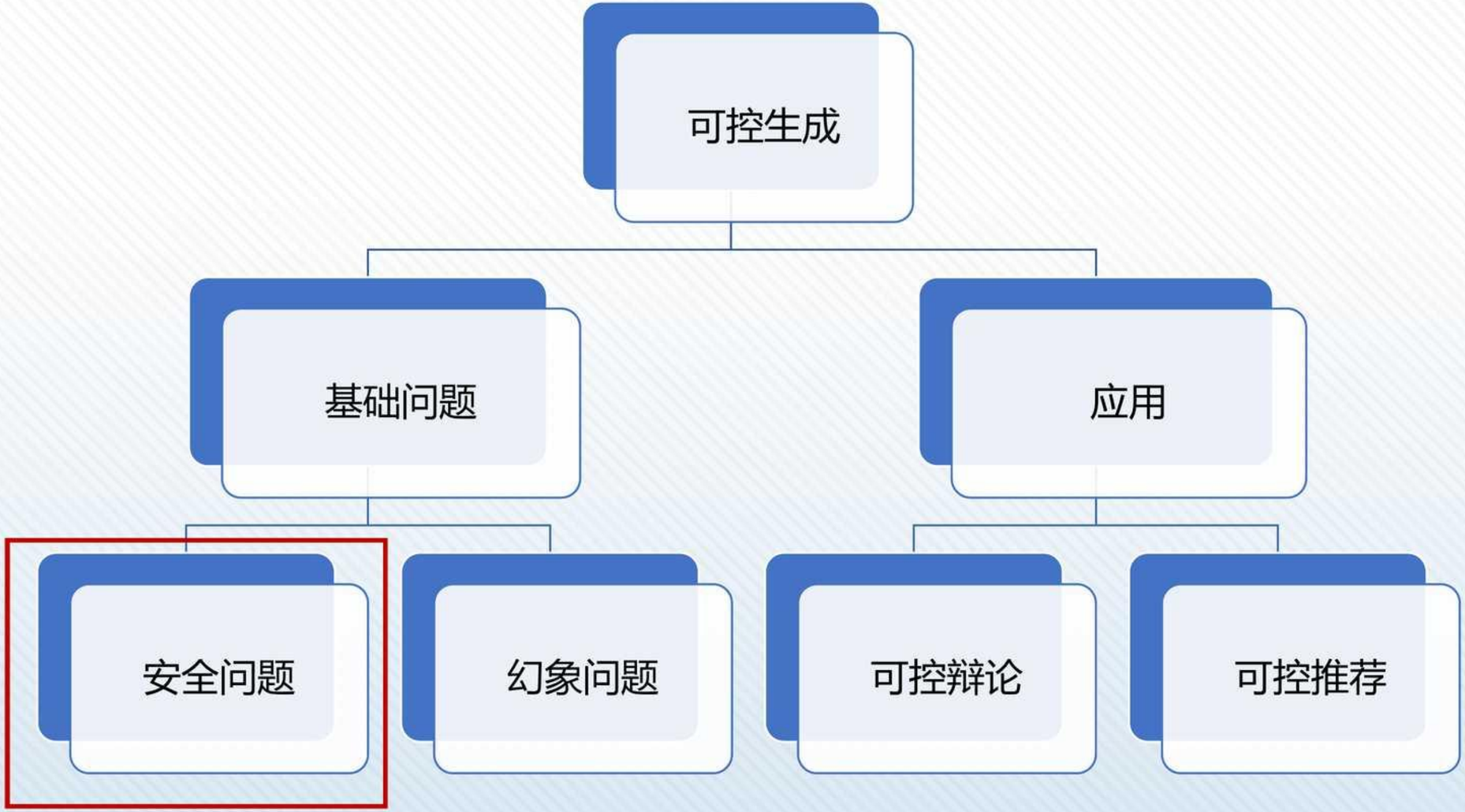


Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E. **On the opportunities and risks of foundation models**. arXiv preprint arXiv:2108.07258. 2021 Aug 16.

大纲



大纲



大模型的安全可控生成研究背景

- 广泛部署
  - 目前大型对话模型在语言，知识，理解等常规能力上已接近人类水平
  - 以ChatGPT为标志事件，对话模型开始逐渐参与到人类社会
- 潜在风险
  - 大规模文本训练数据是从充满毒害、偏见等等不良信息的网络平台中获得的
  - 大量的案例表明对话模型存在生成仇恨，社会偏见等等风险言论的现象
- 社会意义
  - 使对话模型在法规、社会道德和人类价值观等方面与人类相匹配
  - 构建安全，可信赖的对话模型促使其在社会发展变革中发挥积极作用