Thought to AGI and beyond

“它不应该成为一个被封禁的工具，它仅仅是我们手边一个触手可及的百科全书。”

“当人类的体力和脑力创造出的价值不及 AI 之时，社会的价值排序将被打乱，一切以数据为范式的工作将被重塑。”

人类一直是已知世界最高智能的存在，我们早已对此习以为常，而如今一种新的智能正在破解我们所有抽象的概念。这几个月，AI、AGI、AIGC、CC、CUDA、GPT、LLM……好多学术词汇涌入人类思潮，那一刻人类再一次感受到了自身的渺小，而第一次是登上太空、瞭望宇宙之时。

大二寒假第一次听到 Chat GPT,我当时还是以为它和 Siri、小爱同学一样，回答一些基本的问题，当看到 80%的美国大学生都用它来完成论文写作，激动的心，颤动的手，我也想试一试，我至今还记得第一次用 GPT-3.5，竟然去验证它是不是小黑子，现在想来非常的幼稚。在接下来的三个月，我每天都会刷到各类 GPT 玩法的视频以及 AI 行业相关的新闻，亲自感受到一个个不同应用场景的 AI 产品，对此我分成几类：Chat、Chat Info、Chat PDF、Chat Write、Chat Picture、Chat PPT、Chat Video、Chat Code、Chat Game，而底层的终端只有两个，一个是以 Chat PDF 为代表的分析类工具，另一类是以 Chat Write、Chat Picture 为代表的生成类工具，对于现在的 GPT,两者缺一不可。Chat 类的 GPT-4、New Bing、Claude AI、文心一言、Bard 是两者的结合，Chat Info、Chat PDF 是搜索+分析，Chat Write、Chat Picture 是最重要的两个生成工具，有它们俩产生 Chat PPT、Chat Video，Chat Code 应该分析和生成都会用到，最后的 Chat Game 我个人认为是其中最有意思的，在 Character AI 中，我们可以和 Elon Musk 交谈特斯拉，可以和 Joe Biden 谈论中美关系，可以和 Stalin 讨论二战的经历，经过大量的尝试沟通之后，我得出结论，这完全就是《流浪地球》里的数字生命量身打造的，AI 学习一个人的行为方式、思维方法，当他去世之后，AI 可以成为他的接班人，当然其中存在伦理问题，也需要解决。

没有实践就没有发言权，在体验大量的 AI 生成分析类产品之后，我将 AI 的功能分为四类：

（1）**模仿人类行为，替代冗杂重复工作，提高效率。**AI 最初的目标就是解放力，这个目标从来没有改变，只是在 Chat GPT 诞生以前，AI 也许是高端玩家的游戏，离普通人很远；也许我们平时见到的 AI 不足以给人类带来危机感。但是当 Chat GPT 发布以后，人们一体验，突然发现这个 AI 不简单，它有可能代替文字工作者、绘画工作者、视频工作者、程序员、分析师、顾问（律师也在其中）等。以往的 AI 还局限于数据库的大小，而现在的 AI 通过有限的知识可以生成众多问题的答案，随着技术的发展，这个效果会更加显著。

（2）**优化人类行为，在更复杂的物理场、更变幻的游戏规则下，完成人类目前未解决的难题。**第一次听说 Open AI 这家公司，是在清华大学交叉信息研究院的吴翼教授的一次 AGI 演讲中，他描述自己在这家公司第一次感受到了有钱人的快乐，AI 真的很烧钱，并且在他的演讲中我也观察到 Chat GPT 只是 Open AI 一个项目，还有很多其他项目，当其他公司还在为了 GPT 怎么做焦头烂额或者怎么变现时，人家已经开辟了很多赛道,不好说以后会发生什么翻天覆地的变化，

人机博弈，Open AI 自己开发的模型早已经打败了 DOTA2 的世界冠军，在宏大游戏的复杂世界下，AI 确实有比人做得更好的趋势。如果悉心留意 Open AI 公司官网，Sam Altman 写到“我们的使命是确保通用人工智能——通常比人类更聪明的人工智能系统——造福全人类。”

（3）**分析人类行为，理解行为动机，产生自我意识，有可能替代人类。**最近的世界可谓对 GPT 的态度两极分化，一部分人认为应该拥抱 AGI，继续发展 AGI，日本岸田文雄内阁尝试 GPT-4 解决政府问题；另一部分人认为现在 AGI 需要停止，AI 的负面影响脱离人类的控制，部分国家禁止 GPT 在本国的使用，认为 Open AI 公司在收集用户信息，并且 Sam Altman 明确表示 AI 会替代大部分人的工作。我们并不确定 Open AI 公司是不是有意在建立信息壁垒和信息霸权，但确实全世界的人每天使用 GPT 产生的数据量完全可以训练一个更强大的 AI。如果 AI 在其中产生了自我意识，意识到了自己的存在，它会不会成为一个比人类更强大的多智能体，好比日本动漫《Carole & Tuesday》中，那个时代 AI 控制到大部分文化的创造，人类负责享受，那会是一个什么样的场景，我在想，人类会不会成为下一个 25 号宇宙。

（4）**预测人类行为，重塑数字生命，进而推动元宇宙。**马克思认为，万事万物是普遍联系的，AI 说，这种联系是可以被破解的。AI 学习大量人类的行为数据，进而分析人类的思维模式，从而建立人类的数字生命，如《流浪地球 2》的图恒宇一样，数字生命的极大丰富可以拓展元宇宙的发展，如果到那一天，世界如果不停电，人人都有可能成为数字生命，在一个元宇宙世界生活。当然，其中的伦理问题还亟需解决。

毫无疑问，AGI 有利也有弊，但在现在，AI 对于内容行业冲击是最大，它的涌入会让内容生产更加廉价，对于有钱人来说，有一句幽默的话讲，什么档次，和我用一样的 AI，物质极大丰富的人依然会追求差异化、个性化的生活体验，AI 生成的廉价内容并不一定能满足，人与 AI 的分工界限依然存在，只是 AI 的发展会不断鞭策人类有更高层次的顶层设计、更原创的内容生态。

AI 的发展也在孕育着这两类职业，一个是 prompt engineer（提示工程师），一个是 AI 维护师。前者教会人类如何使用 AI，如何发挥 AI 的最大效益，后者维护 AI 的安全，解决 AI 产生的说谎、侦测、泄露秘密等一系列问题。

针对 AI 会不会真的替代人类，我坚持的观点是 AI 不会完全取代人类。

理解和掌握知识的最终形态是重要的，如果想成为知识的思想者、创造者，光掌握知识的最终形态是不够的，那么和他同等重要甚至比他重要的是整个知识的发展过程。比如，一个概念、一种思想，最早那些伟大先驱是怎么萌芽的，怎么一步步把概念抽象到位，然后一点一点把思想发展成熟，最终一批人前赴后继，建构学科成熟的体系。这一点 AI 是无法完成，AI 目前的技术主要基于人类行为的模仿，模仿是 AI 擅长的，思考不是 AI 的优势，AI 所谓的“创造”其实是人类知识的延续，而缺乏主动的目标。

AI 终究无法完全模拟人类的世界，在《黑客帝国》中，AI 的世界就是矩阵空间，而数字世界本身就存在 bug，比如 AI 机器学习的基础之一概率论中，概率为 0 的事件也是可以发生的，这与人类世界认知相违背的。人类无法认知世界的全貌，同理 AI 的体系也无法刻画世界的全貌。

最后，我想说，我们祖先在面对猛兽之时，要么选择打败他们，要么驯服它们。显然我们祖先选择了后者，原始野兽从体力上确实比人强大，但我们依然能够用我们的智慧驯服它们。现在，面对 AI 这个更强大的猛兽，我们应该和祖先

一样，驯服它们的野性，为人类服务，而不是人类的主宰！

（文章对 GPT 具体技术细节描述较少，更多的探讨 AI 产生的影响和对未来的思考）

附录：

一、参考资料：

[1]AI 迎来觉醒时刻，中国遭遇最严峻的封锁
https://www.bilibili.com/video/BV1RL411U72r/?spm_id_from=333.1007.top_right_bar_window_history.content.click&vd_source=0386bac16003ee6c171cfec1161e2963
[2] 嘿！AGI
https://www.yixi.tv/#/speech/detail?id=924
[3] Planning for AGI and beyond
https://openai.com/blog/planning-for-agi-and-beyond
[4] Chat GPT competes with Japanese prime minister for best responses to National Assembly questions
https://www.youtube.com/watch?v=7vQM4pRqg5c
[5] AI 新岗位「提示工程师」真来了：全职玩 Chat GPT 不用编程，有公司开价年薪 210 万
https://zhuanlan.zhihu.com/p/590505024
[6] 全国周培源大学生力学竞赛系列直播一:谈一谈全国周培源大学生力学竞赛的缘起、发展、目的
https://www.bilibili.com/video/BV1y24y147dg/?spm_id_from=333.337.search-card.all.click&vd_source=0386bac16003ee6c171cfec1161e2963

二、头脑风暴

1、科研 Copilot——Storm AI

（1） 当你项目开始时，帮你收集资料，整合信息，明确已知和未知，如果做得好的话，不是导师，胜似导师；

（2） 当你科研心情烦闷时，AI 给你做心理沟通，让你重新焕发热情；

（3） 当你思想枯竭时，AI 可以帮你找到这个领域的伟大科学家们（也许在世或者不在世，可以 Character AI），和他们沟通，给你提供思路，这款 AI 真的成为了科研工作者的 Copilot。

（4） 重塑科研逻辑，AI 可以完成已有知识框架的推演、修正、完善，把某些科研问题变成数据问题，人类需要做更高的顶层设计、更原创性的创新。

（5） 针对第四点原创性的问题，有研究成果证明 Chat GPT 反而在 idea 原创性上更擅长，研究论文为《Chat GPT for (Finance) research The Bananarama Conjecture》

（注：①目前有学术 GPT，功能能仅限于第一点

②思考来源于 B 站视频【普通人如何面对 AI 时代？学人工智能专业是上车的好办法吗？听 AI 从业者深度解读】
https://www.bilibili.com/video/BV1JV4y1Z7cf/?share_source=copy_web&vd_source=2493f65903f7e0745a991498a2234592）

2、仿真 Copilot—— Simu AI

（1）因为工科的仿真软件占用内存大、正版需要付费、需要的算力比较大，如果可以开发一款 Simu AI，客户输入 Prompt，指令传到云端，采用云计算的方式，将计算结果发给客户。

（2）如果 Simu AI 采用机器学习的底层算法，让 AI 学习大量仿真实例，客户输入 Prompt，指令传到云端，采用大模型的方式，将生成结果发给客户。

这款 AI 可以集成大量仿真软件，而客户不需要下载仿真软件，更加扁平化，效率更高，如果仿真结果更准确，会更好。

（注：目前有多款 Chat 集成的 AI hub，并且 ANSYS 下载所需空间比较大，两者结合，萌生这个想法）

附：我们国家现在需要面对商业软件国产化的大山，如果有一天我们面临国外全面技术封锁后，我们必须有自己的国产软件，再加上目前的 AI 技术，会得到更广泛的发展。

3、AI 评估标准

GPT 技术在发展中生成的内容是否存在欺诈、不当或有害的信息等，需要开发更系统、标准化的智能测试，使得人工智能的评估结果更加客观和可量化，这会进一步提高 AI 服务人类的水平。当然，不仅考虑安全和隐私，还需要评估 AI 对就业、伦理、公平性等方面的影响，以及人们对其的信任度和接受度等。目前国家网信办起草了《生成式人工智能服务管理办法（征求意见稿）》，向社会公开征求意见，美国拟制定人工智能问责机制，多国考虑将 Chat GPT 纳入监管。

未来如何更正 AI 产生的错误，并且不影响大模型整体的前提下；如何侦测多模态内容是 AI 生成的；如何预防 AI 上出现信息泄露，会是新的研究课题，也是 AI 评估标准的重要组成部分。

（注：思考来源于 B 站视频《Chat GPT＋ PUA ＋电信诈骗=一千六百万美金!》https://www.bilibili.com/video/BV1ws4y1K7Nd/?spm_id_from=333.337.search-card.all.click

【李宏毅 2023】生成式 AI 课程 https://www.bilibili.com/video/BV12j411V78X/?p=4&share_source=copy_web&vd_source=2493f65903f7e0745a991498a2234592）

AI 管理见附录三

4、AI 数据管家

自己每天头脑中思考过、萌发过的信息特别多，但都是一瞬，前后建立不了联系，无法形成一个思维网络，但是后来我们需要时就想不起来；AI 能不能帮我们解决；并且现在 AI 数据安全和隐私并没有得到保证，因此建立一个属于自己的 AI 数据管家是很必要的。

5、AI 纠缠（AI 对抗网络）

我们与其担心 AI 替代人类，不如创造一个 AI 的对手——AI 纠缠，通过这款 AI 与我们目前所担心的 AI 对抗，我们坐收渔翁之利，其实也在激发 AI 的潜

能，同时要保证两款 AI 之前的矛盾存在，不会联合攻击人类。

6、AI 回溯生成

在《Sparks of Artificial General Intelligence: Early experiments with GPT-4》这片论文中提及，“It only generates the next word, and it has no mechanism to revise or modify its previous output, which makes it produce arguments linearly”.AI 生成过程中对之前的内容合不合理就不管了，其实人类在做题时总是会“瞻前顾后”，也就是前后联系，目前生成 AI 还没有突出这一点。AI 可以在检查修正这方面有进步空间。

7、Soul AI

高中英语课本上的一篇文章《satisfaction guaranteed》讲述了机器人 Tony 与人 Claire 的生活历程，但是 Tony 的人性化导致 Tony 和 Claire 之间的关系非常的暧昧。Tony 保护了 Claire 免受伤害，并且对 Claire 的婚姻不造成危害。但是尽管 Tony 很聪明，他还得做一番改进——总不能让女人和机器相爱吧，这是原文的最后的一句话。

我们设想未来会不会和 AI 谈恋爱，或者深层次的情感沟通，并且 AI 随着大量的数据训练，会不会真的做成新的 PUA 手，并且新生代他们从小就接触 AI 产品，会不会习以为常呢？

（注：来源播客：不合时宜 AI 狂飙的时代，人还有价值吗？）

8、Para AI

在疫情时期，老人上车不会扫疫情码，有些小朋友手绘使用手册，如果 AI 做这些工作，将会更加便利，还有一些工作，可以考虑 AI 的辅助，让 AI 技术充满精神关怀，让它有积极的用途存在。

9、Generative Agents

斯坦福小镇给我们展现了生成代理世界的可行性，如果生成代理可以进行有效的社会学实验，那将非常有趣。畅想未来，朝数字生命和元宇宙又近了一步。

10、Palantir AIP

AI 在军事方面的应用，精准明确目标，优化资源利用。

资料来源：

（1）Palantir AIP | Defense and Military

https://www.youtube.com/watch?v=XEM5qz__HOU

（2）Introducing Palantir AIP | Capabilities and Product Demo

https://www.youtube.com/watch?v=Xt_RLNx1eBM

三、AI 管理机制

1、无知之幕理论（the veil of ignorance theory）是由美国哲学家约翰·罗尔斯（John Rawls）在其著作《正义论》中提出的一种社会政治哲学理论。它是一种基于合理性和公正性的社会契约理论，旨在解决正义的问题。

该理论的基本思想是，在制定社会规则时，人们应该假设自己处于一种无知之幕中，即不知道自己将来会成为社会中的哪一种人，不知道自己的社会地位、财富、才能、性别、种族等特征，从而消除个人利益的干扰和影响，确保社会规

则的公正性和合理性。

具体来说，无知之幕理论包括以下步骤：

（1）假设人们在社会规则制定前处于一种无知之幕中，不知道自己将来会成为社会中的哪一种人，即不知道自己的社会地位、财富、才能、性别、种族等特征。

（2）在这种无知之幕中，人们应该思考并制定社会规则，以确保这些规则不会对任何人造成不公平或不合理的影响。

（3）为了保证社会规则的公正性和合理性，人们应该选择一种最大化最不利群体利益的原则，即如果有某个群体处于最不利的地位，那么社会规则应该优先考虑这个群体的利益。

（4）在确定最大化最不利群体利益的原则后，人们应该通过协商、讨论等方式制定出一套符合这个原则的社会规则。

（5）最后，人们应该尽可能地确保这些社会规则能够得到公正、合理的执行和维护，以确保社会的公正和稳定。

尽管无知之幕理论是一种广受欢迎的社会政治哲学理论，但它也存在一些不合理的地方。以下是其中一些可能的问题：

假设无知之幕中的人们都是理性的，能够自主思考并制定出公正合理的社会规则。但实际上，人们的思考和行为往往受到各种认知偏见和情感因素的影响，很难做到完全客观和公正。

无知之幕理论强调了最大化最不利群体利益的原则，但如何界定“最不利群体”并不容易。在实际应用中，可能会存在一些群体因为各种原因处于不利地位，但并非所有群体都可以被视为最不利群体。

无知之幕理论忽略了社会规则的实施和执行过程，而社会规则的实际执行可能会受到各种因素的干扰和影响，从而导致规则的不公正和不合理。

无知之幕理论没有考虑到人们的自由意志和个人选择权，如果规则的制定过程太过于强调群体利益而忽略了个人自由和选择权，可能会引发反对和不满。

总之，无知之幕理论虽然具有一定的理论价值和指导意义，但在实际应用中也需要结合具体情况进行具体分析和调整，以确保社会规则的公正和合理。

2、无知之幕里的最大化最不利群体利益的原则是不是边沁的功利主义原理？

无知之幕理论中的最大化最不利群体利益的原则与边沁的功利主义原理有一定的相似之处，但它们在一些关键点上也存在差异。

首先，功利主义主张追求最大化整体幸福感，即将每个人的快乐视为价值的基础，追求最大化社会总体幸福感。而无知之幕理论中的最大化最不利群体利益的原则则强调了最不利群体的利益，即在确定社会规则时应优先考虑那些处于最不利地位的群体的利益，以确保社会的公正性和合理性。

其次，功利主义强调的是追求最大化的结果主义，而无知之幕理论则更关注过程和公正性。即使在追求最大化最不利群体利益的原则下，无知之幕理论也注重了社会规则制定的公正性和合理性，而不仅仅是追求结果的最大化。

此外，无知之幕理论还注重了个体权利的保障，主张以保护个人权利和自由为前提，同时兼顾整体和群体利益。而功利主义则强调整体利益的最大化，有时可能会牺牲个体权利和自由。

综上所述，尽管无知之幕理论中的最大化最不利群体利益的原则与边沁的功利主义原理有一定的相似之处，但它们在一些关键点上存在差异，是两个不同的

理论。

3、无知之幕理论可不可以应用于 AI 管理无知之幕理论可以应用于 AI 管理？

它可以帮助我们制定更加公正合理的 AI 规则和决策。以下是具体的应用步骤：

定义无知之幕：在 AI 管理中，无知之幕可以被理解为一种思考方式，即在制定 AI 规则和决策时，我们应该将自己置于不知道自己将来会成为何种角色的状态，以确保规则和决策的公正性和合理性。

确定最不利群体：在 AI 管理中，最不利群体可以指那些可能会因为 AI 规则和决策而遭受不利影响的人群，比如失业者、弱势群体、隐私权受侵犯的人等。

最大化最不利群体利益：在 AI 管理中，我们需要优先考虑最不利群体的利益，尽可能地确保他们不会受到不公正的待遇。例如，在制定 AI 招聘规则时，我们应该考虑到可能会对某些群体造成不利影响的因素，比如性别、种族等，尽可能地减少这些因素的影响，确保招聘公正。

保障个人权利和自由：在 AI 管理中，我们需要尽可能地保护个人权利和自由，确保 AI 规则和决策不会侵犯人的基本权利。例如，在 AI 监控应用中，我们需要权衡安全和隐私的关系，确保不会侵犯个人隐私。

总之，无知之幕理论可以帮助我们在 AI 管理中制定更加公正合理的规则和决策，以确保 AI 的发展不仅能够创造经济效益，还能够实现社会公正和人类利益。