Reframing Superintelligence:
The comprehensive AI services model

Eric Drexler
Senior Research Fellow, AI Macrostrategy
Future of Humanity Institute
University of Oxford

OpenAI
21 June 2018
UC Berkeley

Where do AI systems come from?

— Research and development

What do AI systems do?

— Provide services within budgeted time and resources

How do we get general, superintelligent-level AI?

— Asymptotically recursive AI R&D automation

⇒ Comprehensive, high-level AI services (“CAIS”)

>>> *CAIS includes developing and managing new services* <<<

Think of tiling task-space with AI capabilities

Fundamental considerations:

- Learning and competence are distinct, separable aspects of intelligence.
- Products of general learning algorithms can have stable competencies.
- Broad knowledge and reasoning can support focused, episodic tasks.
- Predictive modeling of human concerns & approval is a non-agent service.
- Complex tasks can build on stable, tested, bounded competencies.
- Large task spaces can be tiled with narrower competencies.

Embedding spaces can match tasks with services

— Services have properties (kinds, specialties → spaces, embeddings)
— Tasks have properties (kinds, instances → spaces, embeddings)
— *Co-embedding* enables service selection by neighbor lookup

Cosine similarity "lookup" is pervasive in deep learning

— Attention over memory by cosine similarity (if apply thresholds)
— Single and few-shot learning schemes (kNN in embedding spaces)
— Gated MoE models ("Outrageously large neural networks")
— *With weight and input normalization,* ReLU(matrix · vector)

— Neighbor search works at internet scale —

"Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs" (2016)

"Fast Approximate Nearest Neighbor Search With The Navigating Spreading-out Graph" (2018)

Some points of overlap with Paul's proposals:

— Capabilities increase incrementally under direct and indirect supervision.

— Multiple system can provide checks and balances on one another.

— AI systems are used to develop other AI systems (not "self" transformation).

— Aggregate human and AI capabilities always exceed those of new systems.

— Systems can perform bounded tasks at a potentially superintelligent level.

Key points of conflict with Paul's proposals:

— ????

… CAIS models seem more complementary than conflicting:
Paul's proposed models align services with human goals

Some questionable steps in typical AGI-agent scenarios:

1) Advanced AI capabilities arrive in the form of an individual system that is conceptually a black box *(~Paul: Always ask what happens first).*

2) The intelligent capabilities play a mind-like role in an agent concerned with future states of the world.

3) The AI system gains superintelligent capabilities through radical self-directed transformation.

4) The AI agent pursues goals that will by default be served by seeking greater capability and autonomy *(& might take over the world on the first try).*

5) There is no *accessible, attractive, and less-risky* approach to satisfying human demand for general-purpose AI capabilities.

What seems clear:

— Recursive improvement is not about self-improving agents.

— Learning must be distinguished from performance.

— On-demand service development supports generality.

— Models of AI systems should not *default* to agency.

— Advanced agents emerge *in the context* of advanced services.

— Can’t assume an overwhelming drive toward risky AGI.

— ∴ Need to review research assumptions, agendas in a ~CAIS frame

Home page: **goo.gl/1HEbcH**

The R&D automation model of recursive improvement	**(2 min read)**
To understand AI prospects, focus on AI *services,* not AI *systems*	**(6 min read)**
Competitive AI capabilities will not be boxed	**(5 min read)**
R&D automation is compatible with both strong and weak centralization	**(2 min read)**
Standard definitions of "superintelligence" conflate learning with competence	**(3 min read)**
Rational-agent models place intelligence in an implicitly anthropomorphic frame	**(4 min read)**
Broadly-competent systems coordinate narrower competencies	**(4 min read)**
Development-oriented models align with deeply-structured AI systems	**(4 min read)**
Comprehensive AI services neither require nor entail classic AGI	**(7 min read)**
AGI implementation technologies would enable non-AGI alternatives	**(4 min read)**
Opaque algorithms are compatible with functional transparency and control	**(4 min read)**
End-to-end reinforcement learning is compatible with the AI-services model	**(5 min read)**
Aggregated experience and centralized learning support AI-agent applications	**(5 min read)**
Could 1 PFLOP/s systems exceed the basic functional capacity of the brain?	**(7 min read)**
How do neural and symbolic technologies mesh?	**(14 min read)**

Home page: goo.gl/1HEbcH

The R&D automation model of recursive improvement	**(2 min read)**
AGI agents should be considered in an AI-services context	**(5 min read)**
The orthogonality thesis undercuts the generality of instrumental convergence	**(6 min read)**
Strong optimization can increase AI safety	**(5 min read)**
Broad world knowledge is compatible with strong (and safe) task focus	**(6 min read)**
Machine learning can develop predictive models of human approval	**(4 min read)**
Human oversight need not impede fast, recursive AI technology improvement	**(4 min read)**
Advanced AI development systems can support effective human guidance	**(6 min read)**
Optimized advice need not be optimized to induce its acceptance	**(5 min read)**
Superintelligent-level systems can safely provide design and planning services	**(7 min read)**
Collusion among superintelligent oracles can readily be avoided	**(5 min read)**
Competitive pressures provide little incentive to transfer strategic choices to AI systems	**(3 min read)**
Supercapabilities do not entail superpowers	**(2 min read)**
Unaligned superintelligent agents need not threaten world stability	**(5 min read)**
Predictable aspects of future knowledge are key to AI safety planning	**(4 min read)**
The AI-services model has both descriptive and prescriptive aspects	**(2 min read)**
Desiderata and directions for interim AI safety guidelines	**(5 min read)**

Characteristics	***...in AGI-agent models***	***...in CAIS models***
Recursive improvement:	Self-improving agent	Automated AI R&D
Stronger optimization:	More powerful, general	More effective, focused
Takeoff threshold:	Agent capabilities	Technological capabilities
Takeoff dynamics:	Autonomous, abrupt	Managed, potentially swift
Developer experience:	Unprecedented, novel	Extensive, incremental
Typical embodiments:	AI agents	Devices, cloud services
Autonomous change:	Self transformation	Task-focused learning
Action evaluation:	Greater global utility	Better output metrics
Default goals:	Resources, autonomy	Task completion
Getting results:	Request AGI action	Employ a service
Autonomous AGI:	Uniquely valuable	Needlessly risky
Primary risks:	Autonomous AI	Human bad actors