Reframing Superintelligence

Comprehensive AI Services as General Intelligence

K. Eric Drexler · FHI Technical Report #2019-1

Extended table of contents

Abstract

Preface

§I Introduction: From R&D automation to comprehensive AI Services

Thesis. Responsible development of AI technologies can provide an increasingly comprehensive range of superintelligent-level AI services—including the service of developing new services—and can thereby deliver the value of general-purpose AI while avoiding the risks associated with self-modifying AI agents.

§I.1 Summary
§I.2 The trajectory of AI development reframes AI prospects
§I.3 R&D automation suggests a technology-centered model of recursive improvement
§I.4 R&D automation suggests a service-centered model of general intelligence
§I.5 The services model abstracts functionality from implementation
§I.6 The R&D automation model distinguishes development from functionality
§I.7 Language translation exemplifies a safe, potentially superintelligent service
§I.8 Predictive models of human (dis)approval can aid AI goal alignment
§I.9 The R&D-automation/CAIS model reframes prospects for superintelligence

§II Overview: Questions, propositions, and topics

§II.1 Summary
§II.2 Reframing prospects for an intelligence explosion
- Does recursive improvement imply self-transforming agents?
- Would self-transforming agents provide uniquely valuable functionality?
- Can fast recursive improvement be controlled and managed?
- Would general learning algorithms produce systems with general competence?

§II.3 Reframing the nature of advanced machine intelligence
- Is human learning an appropriate model for AI development?
- Does stronger optimization imply greater capability?
- Do broad knowledge and deep world models imply broad AI capabilities?
- Must we model SI-level systems as rational, utility-maximizing agents?
- Must we model SI-level systems as unitary and opaque?

§II.4 Reframing the relationship between goals and intelligence
- What does the orthogonality thesis imply for the generality of convergent instrumental goals?
- How broad is the basin of attraction for convergent instrumental goals?

§II.5 Reframing the problem of using and controlling advanced AI
- Would the ability to implement potentially-risky self-transforming agents strongly motivate their development?
- How can human learning during AI development contribute to current studies of AI safety strategies?
- Can we avoid strong trade-offs between development speed and human oversight?

- Can we architect safe, superintelligent-level design and planning services?
- Will competitive pressures force decision-makers to transfer strategic decisions to AI systems?
- What does the R&D-automation/AI-services model imply for studies of conventional vs. extreme AI-risk concerns?
- What can agent-oriented studies of AI safety contribute, if risky AI agents are optional?

§II.6 Reframing near- and long-term considerations in AI safety and strategy

- Could widely available current or near-term hardware support superintelligence?
- What kinds of near-term safety-oriented guidelines might be feasible and useful?
- How can near-term differential technology development contribute to safety?
- In the long term, are unaligned superintelligent agents compatible with safety?

§II.7 Conclusions

§1 R&D automation provides the most direct path to an intelligence explosion

Thesis. The most direct path to accelerating AI-enabled progress in AI technology leads through AI R&D automation, not through self-transforming AI agents.

§1.1 Summary

§1.2 AI-enabled AI development could lead to an intelligence explosion

§1.3 Risk models have envisioned AGI agents driving an intelligence explosion

§1.4 Self-transforming AI agents have no natural role in recursive improvement

§1.5 The direct path to an intelligence explosion does not rely on AGI agents

§2 Standard definitions of “superintelligence” conflate learning with competence

Thesis. By implicitly conflating learning with competence, standard definitions of “superintelligence” fail to capture what we mean by “intelligence” and obscure much of the potential solution space for AI control problems.

§2.1 Summary

§2.2 Superintelligence has been defined in terms of adult human competence

§2.3 “Intelligence” often refers instead to learning capacity

§2.4 Learning and competence are separable in principle and practice

§2.5 Patterns of AI learning and competence differ radically from humans’

§2.6 Distinguishing learning from competence is crucial to understanding potential AI control strategies

§3 To understand AI prospects, focus on services, not implementations

Thesis. Shifting the focus of attention from AI systems to AI services yields a more uniform and tractable model of artificial general intelligence that reframes the problem of aligning superintelligent-level AI functionality with human purposes.

§3.1 Summary

§3.2 The instrumental function of AI technologies is to provide services

§3.3 General intelligence is equivalent to general capability development

§3.4 The ability to learn is a capability

§3.5 Implementing new capabilities does not require “self modification”

§3.6 Service-centered models highlight differentiated, task-focused functionality

§3.7 Service-centered models harmonize with practice in software engineering

§3.8 Service-centered AI architectures can facilitate AI alignment

§4 The AI-services model includes both descriptive and prescriptive aspects

Thesis. The AI-services model both *describes* the architecture of current and prospective high-level AI applications, and *prescribes* patterns of development that can foster safety without impeding the speed and efficiency of AI development.

§4.1 Summary
§4.2 The AI-services model describes current AI development
§4.3 AI-service development scales to comprehensive, SI-level services
§4.4 Adherence to the AI-services model aligns with AI safety
§4.5 Adherence to the AI-services model seems desirable, natural, and practical

§5 Rational-agent models place intelligence in an implicitly anthropomorphic frame

Thesis. It is a mistake to frame intelligence as a property of mind-like systems, whether these systems are overtly anthropomorphic or abstracted into decision-making processes that guide rational agents.

§5.1 Summary
§5.2 The concept of mind has framed our concept of intelligence
§5.3 Studies of advanced AI often posit intelligence in a psychomorphic role
§5.4 Intelligent systems need not be psychomorphic
§5.5 Engineering and biological evolution differ profoundly
§5.6 Studies of AI prospects have often made tacitly biological assumptions
§5.7 Potential mind-like systems are situated in a more general space of potential intelligent systems

§6 A system of AI services is not equivalent to a utility maximizing agent

Thesis. The conditions for von Neumann–Morgenstern rationality do not imply that *systems composed of AI services* will act as utility-maximizing agents, hence the design space for manageable superintelligent-level systems is broader than often supposed.

§6.1 Summary
§6.2 Systems of SI-level agents have been assumed to act as a single agent
§6.3 Individual service providers can be modeled as individual agents
§6.4 Trivial agents can readily satisfy the conditions for VNM rationality
§6.5 Trivially-rational agents can employ reasoning capacity of any scope
§6.6 High intelligence does not imply optimization of broad utility functions
§6.7 Systems composed of rational agents need not maximize a utility function
§6.8 Multi-agent systems are structurally inequivalent to single agents
§6.9 Problematic AI services need not be problematic AGI agents
§6.10 The AI-services model expands the solution-space for addressing AI risks

§7 Training agents in human-like environments can provide useful, bounded services

Thesis. Training agents with curiosity and imagination to perform human-like tasks in human-like environments can yield systems that provide bounded, human-like services.

§7.1 Summary
§7.2 Does training on human-like tasks conflict with the AI-services model?
§7.3 Human-like learning may be essential to developing general intelligence
§7.4 Current methods build curious, imaginative agents
§7.5 Human-like competencies do not imply human-like goal structures

§7.6 Open-ended learning can develop skills applicable to bounded tasks
§7.7 Human-like world-oriented learning nonetheless brings unique risks

§8 Strong optimization can strongly constrain AI capabilities, behavior, and effects

Thesis. Strong (even superintelligent-level) optimization can be applied to increase AI safety by strongly constraining the capabilities, behavior, and effects of AI systems.

§8.1 Summary
§8.2 Strong optimization power need not increase AI capability and risk
§8.3 Strong optimization is a strong constraint
§8.4 Optimization of AI systems can reduce unintended consequences
§8.5 Strong external optimization can strongly constrain internal capabilities
§8.6 Optimizing an AI system for a bounded task is itself a bounded task
§8.7 Superintelligent-level optimization can contribute to AI safety

§9 Opaque algorithms are compatible with functional transparency and control

Thesis. Although transparency is desirable, opacity at the level of algorithms and representations need not greatly impair understanding of AI systems at higher levels of functionality.

§9.1 Summary
§9.2 Deep-learning methods are opaque and may remain so
§9.3 The scope of information and competencies can be fuzzy, yet bounded
§9.4 Restricting resources and information at boundaries constrains capabilities
§9.5 Providing external capabilities can constrain internal capabilities
§9.6 Deep learning can help interpret internal representations
§9.7 Task-space models can enable a kind of “mind reading”
§9.8 The application of control measures can be adapted to experience and circumstances

§10 R&D automation dissociates recursive improvement from AI agency

Thesis. In automation of AI research and development, AI agents are useful products, not necessary components.

§10.1 Summary
§10.2 R&D automation can employ on diverse, specialized AI tools
§10.3 AI R&D automation will reflect universal aspects of R&D processes
§10.4 AI R&D automation will reflect the structure of AI development tasks
§10.5 AI R&D automation leads toward recursive technology improvement
§10.6 General SI-level functionality does not require general SI-level agents
§10.7 The R&D-automation model reframes the role of AI safety studies and offers potential affordances for addressing AI safety problems

§11 Potential AGI-enabling technologies also enable comprehensive AI services

Thesis. If future AI technologies can implement self-transforming AGI agents, then similar capabilities could more easily be applied to implement open, comprehensive AI services.

§11.1 Summary
§11.2 In runaway-AGI scenarios, self-improvement precedes risky competencies
§11.3 “Self”-improvement mechanisms would first accelerate R&D

§11.4 “Self”-improvement mechanisms have no special connection to agents
§11.5 Transparency and control need not impede the pace of AI development
§11.6 Optimization pressures sharpen task focus
§11.7 Problematic emergent behaviors differ from classic AGI risks
§11.8 Potential AGI technologies might best be applied to automate development of comprehensive AI services

§12 AGI agents offer no compelling value

Thesis. Because general AI-development capabilities can provide stable, comprehensive AI services, there is no *compelling, practical motivation* for undertaking the more difficult and potentially risky implementation of self-modifying AGI agents.

§12.1 Summary
§12.2 Would AGI development deliver compelling value?
§12.3 AI systems deliver value by delivering services
§12.4 Providing diverse AI services calls for diverse AI capabilities
§12.5 Expanding AI-application services calls for AI-development services
§12.6 The AGI and CAIS models organize similar functions in different ways
§12.7 The CAIS model provides additional safety-relevant affordances
§12.8 The CAIS model enables competition and adversarial checks
§12.9 The CAIS model offers generic advantages over classic AGI models
§12.10 CAIS affordances mitigate but do not solve AGI-control problems

§13 AGI-agent models entail greater complexity than CAIS

Thesis. Relative to comprehensive AI services (CAIS), and contrary to widespread intuitions, the classic AGI-agent model implicitly increases (while obscuring) the complexity and challenges of self-improvement, general functionality, and AI goal alignment.

§13.1 Summary
§13.2 Classic AGI models neither simplify nor explain self improvement
§13.3 Classic AGI models neither simplify nor explain general AI capabilities
§13.4 Classic AGI models increase the challenges of AI goal alignment
§13.5 The CAIS model addresses a range of problems without sacrificing efficiency or generality

§14 The AI-services model brings ample risks

Thesis. High-level AI services could facilitate the development or emergence of dangerous agents, empower bad actors, and accelerate the development of seductive AI applications with harmful effects.

§14.1 Summary
§14.2 Prospects for general, high-level AI services reframe AI risks
§14.3 CAIS capabilities could mitigate a range of AGI risks
§14.4 CAIS capabilities could facilitate the development of dangerous AGI agents
§14.5 CAIS capabilities could empower bad actors.
§14.6 CAIS capabilities could facilitate disruptive applications
§14.7 CAIS capabilities could facilitate seductive and addictive applications
§14.8 Conditions for avoiding emergent agent-like behaviors call for further study

§15 Development-oriented models align with deeply-structured AI systems

Thesis. Unitary, unstructured models of superintelligent systems are natural objects of theoretical study, but development-oriented models suggest that advanced AI systems will in practice comprise deeply structured compositions of differentiated components.

§15.1 Summary
§15.2 AI safety research has often focused on unstructured rational-agent models
§15.3 Structured systems are products of structured development
§15.4 AI development naturally produces structured AI systems
§15.5 Structure arises from composing components, not partitioning unitary systems
§15.6 A development-oriented approach to deeply structured systems suggests a broad range of topics for further inquiry

§16 Aggregated experience and centralized learning support AI-agent applications

Thesis. Centralized learning based on aggregated experience has strong advantages over local learning based on individual experience, and will likely dominate the development of advanced AI-agent applications.

§16.1 Summary
§16.2 Discussions often assume learning centered on individual agents
§16.3 Naïve scaling to multi-agent systems replicates individual agents
§16.4 Self-driving vehicles illustrate the power of aggregating experience
§16.5 Efficiently organized machine learning contrasts sharply with human learning
§16.6 Aggregated learning speeds development and amortizes costs
§16.7 The advantages of aggregated, amortized learning have implications for prospective AI-agent applications

§17 End-to-end reinforcement learning is compatible with the AI-services model

Thesis. End-to-end training and reinforcement learning fit naturally within integrated AI-service architectures that exploit differentiated AI components.

§17.1 Summary
§17.2 RL and end-to-end training tend to produce black-box systems
§17.3 RL and end-to-end training are powerful, yet bounded in scope
§17.4 General capabilities comprise many tasks and end-to-end relationships
§17.5 Broad capabilities are best built by composing well-focused competencies
§17.6 Deep RL can contribute to R&D automation within the CAIS model of general AI

§18 Reinforcement learning systems are not equivalent to reward-seeking agents

Thesis. RL systems are (sometimes) used to train agents, but are not themselves agents that seek utility-like RL rewards.

§18.1 Summary
§18.2 Reinforcement learning systems differ sharply from utility-directed agents
§18.3 RL systems are neither trained agents nor RL-system developers
§18.4 RL systems do not seek RL rewards, and need not produce agents
§18.5 RL rewards are not, in general, treated as increments in utility
§18.6 Experience aggregation blurs the concept of individual reward

§18.7 RL algorithms implicitly compete for approval
§18.8 Distinctions between system levels facilitate transparency and control
§18.9 RL-driven systems remain potentially dangerous

§19 The orthogonality thesis undercuts the generality of instrumental convergence

Thesis. If any level of intelligence can be applied to any goal, then superintelligent-level systems can pursue goals for which the pursuit of the classic instrumentally-convergent subgoals would offer no value.

§19.1 Summary
§19.2 The thesis: Any level of intelligence can be applied to any goal (more or less)
§19.3 A wide range of goals will engender convergent instrumental subgoals
§19.4 Not all goals engender IC subgoals
§19.5 Not all intelligent systems are goal-seeking agents in the relevant sense
§19.6 Comprehensive services can be implemented by systems with bounded goals
§19.7 IC goals naturally arise as tropisms and as intended services
§19.8 Systems with tropisms are not equivalent to agents with “will”

§20 Collusion among superintelligent oracles can readily be avoided

Thesis. Because perverse collusion among AI systems would be fragile and readily avoided, there is no obstacle to applying diverse, high-level AI resources to problems of AI safety.

§20.1 Summary
§20.2 Trustworthiness can be an emergent property
§20.3 A range of conditions may facilitate or disrupt collusion
- Conditions that tend to facilitate collusion among actors
- Contrasting conditions that tend to disrupt collusion among actors
- Natural architectures for problem-solving systems

§20.4 Collusion is fragile and easily disrupted

§21 Broad world knowledge can support safe task performance

Thesis. Broad, deep knowledge about the world is compatible with safe, stable, high-level task performance, including applications that support AI safety.

§21.1 Summary
§21.2 Bounding task focus does not require bounding world knowledge
§21.3 Extensive world knowledge can improve (e.g.) translation
§21.4 Current MT systems are trained on open-ended text corpora
§21.5 Current systems develop language-independent representations of meaning
§21.6 Scalable MT approaches could potentially exploit extensive world knowledge
§21.7 Specialized modules can be trained on diverse, overlapping domains
§21.8 Safe task focus is compatible with broad, SI-level world knowledge
§21.9 Strong task focus does not require formal task specification

§22 Machine learning can develop predictive models of human approval

Thesis. By exploiting existing corpora that reflect human responses to actions and events, advanced ML systems could develop predictive models of human approval with potential applications to AI safety.

§22.1 Summary
§22.2 Advanced ML technologies will precede advanced AI agents

§22.3 Advanced ML can implement broad predictive models of human approval
§22.4 Text, video, and crowd-sourced challenges can provide training data
§22.5 Predictive models of human approval can improve AI safety
§22.6 Prospects for approval modeling suggest topics for further inquiry

§23 AI development systems can support effective human guidance

Thesis. Advanced, interactive AI development resources could greatly facilitate the use of human guidance in developing safe, high-level AI services.

§23.1 Summary
§23.2 Facilitating human guidance is part of the AI-application development task
§23.3 Development tasks include task selection, system design, training, testing, deployment, in-use feedback, and upgrades
§23.4 We should assume effective use of natural-language understanding
§23.5 Generic models of human (dis)approval can provide useful priors
§23.6 Bounded task objectives can be described and circumscribed
§23.7 Observation can help systems learn to perform human tasks
§23.8 Deployment at scale enables aggregated experience and centralized learning
§23.9 Recourse to human advice will often be economical and effective
§23.10 AI-enabled criticism and monitoring can strengthen oversight
§23.11 AI-enabled AI development could both accelerate application development and facilitate human guidance

§24 Human oversight need not impede fast, recursive AI technology improvement

Thesis. Human guidance and safety-oriented monitoring can operate outside core technology development loops, and hence are compatible with fast, recursive AI technology improvement.

§24.1 Summary
§24.2 Must pressure to accelerate AI technology development increase risk?
§24.3 Basic technology research differs from world-oriented applications
§24.4 We can distinguish between human participation, guidance, and monitoring
§24.5 Guidance and monitoring can operate outside the central AI R&D loop
§24.6 Fast, asymptotically-recursive basic research need not sacrifice safety
§24.7 World-oriented applications bring a different range of concerns

§25 Optimized advice need not be optimized to induce its acceptance

Thesis. Advice optimized to produce *results* may be manipulative, optimized to induce a client's acceptance; advice optimized to produce results *conditioned on its acceptance* will be neutral in this regard.

§25.1 Summary
§25.2 Background (1): Classic concerns
§25.3 Background (2): Development-oriented models
§25.4 Optimization for results favors manipulating clients' decisions
§25.5 Optimization for results conditioned on actions does not entail optimization to manipulate clients' decisions
§25.6 Oracles can suggest options with projected costs, benefits, and risks
§25.7 Competitive pressures may nonetheless favor AI systems that produce perversely appealing messages

§26 Superintelligent-level systems can safely provide design and planning services

Thesis. Superintelligent-level AI systems can safely converse with humans, perform creative search, and propose designs for systems to be implemented and deployed in the world.

§26.1 Summary
§26.2 Design engineering is a concrete example of a planning task
§26.3 AI-based design systems match classic templates for emergent AI-agent risk
§26.4 High-level design tasks comprise distinct non-agent-like subtasks
§26.5 Real-world task structures favor finer-grained task decomposition
§26.6 Use of task-oriented components minimizes or avoids classic AI risks
§26.7 Effective human oversight is not an impediment, but a source of value
§26.8 SI-level systems could solve more AI-control problems than they create
§26.9 Models of human concerns and (dis)approval can augment direct oversight
§26.10 The pursuit of superintelligent-level AI design services need not entail classic AI-agent risks

§27 Competitive pressures provide little incentive to transfer strategic decisions to AI systems

Thesis. Requirements for swift response may motivate transfer of tactical-level control to AI systems; at a strategic level, however, humans will have both time and good reason to consider suggested alternatives.

§27.1 Summary
§27.2 Pressures for speed and quality can favor AI control of decisions
§27.3 Speed is often critical in selecting and executing “tactical” actions
§27.4 Quality is more important than speed in strategic planning
§27.5 System that can make excellent decisions could suggest excellent options
§27.6 Human choice among strategies does not preclude swift response to change
§27.7 Senior human decision makers will likely choose to retain their authority

§28 Automating biomedical R&D does not require defining human welfare

Thesis. Superintelligent-level support for tasks as broad as biomedical research need not incur the classic problems of AGI value alignment.

§28.1 Summary
§28.2 Broad, unitary tasks could present broad problems of value alignment
§28.3 Diverse AI systems could automate and coordinate diverse research tasks
§28.4 Human oversight can be supported by AI tools
§28.5 Strong task alignment does not require formal task specification
§28.6 The advantages of assigning broad, unitary tasks to AGI agents are questionable

§29 The AI-services model reframes the potential roles of AGI agents

Thesis. Potential AGI agents should be considered in the context of a world that will (or readily could) have prior access to general intelligence in the form of comprehensive AI services.

§29.1 Summary
§29.2 It has been common to envision AGI agents in a weak-AI context
§29.3 Broad, SI-level services will (or readily could) precede SI-level AI agents
§29.4 SI-level services will enable the implementation of AGI agents
§29.5 SI-level advisory and security services could limit AGI-agent risks

§29.6 SI-level capabilities could mitigate tensions between security concerns and ethical treatment of non-human persons

§29.7 Prospects for superintelligence should be considered in the context of an SI-level AI services milieu

§30 Risky AI can help develop safe AI

Thesis. Complex, unitary, untrusted AI systems could be leveraged to produce non-problematic general-learning kernels through competitive optimization for objectives that heavily weight minimizing description length.

§30.1 Summary

§30.2 A familiar threat model posits opaque, self-improving, untrusted AI systems

§30.3 Open-ended “self-improvement” implies strong, general AI implementation capabilities

§30.4 Successful system development can be recapitulated with variations

§30.5 Optimization can favor the production of compact, general learning kernels

§30.6 Competitive optimization for compactness can exclude problematic information and competencies

§30.7 Exclusion of problematic content can provide a safe basis for developing general capabilities

§31 Supercapabilities do not entail “superpowers”

Thesis. By definition, any given AI system can have “cognitive superpowers” only if others do not, hence (strategic) superpowers should be clearly distinguished from (technological) supercapabilities.

§31.1 Summary

§31.2 AI-enabled capabilities could provide decisive strategic advantages

§31.3 Superpowers must not be confused with supercapabilities

§32 Unaligned superintelligent agents need not threaten world stability

Thesis. A well-prepared world, able to deploy extensive, superintelligent-level security resources, need not be vulnerable to subsequent takeover by superintelligent agents.

§32.1 Summary

§32.2 General, SI-level capabilities can precede AGI agents

§32.3 SI-level capabilities could be applied to strengthen defensive stability

§32.4 Unopposed preparation enables strong defensive capabilities

§32.5 Strong defensive capabilities can constrain problematic agents

- 1) Preparatory, SI-level red-team/blue-team design competition
- 2) Anticipatory deployment of well-resourced, SI-level security services
- 3) Ongoing application of effective, physically-oriented surveillance
- 4) Ongoing application of effective, physically-oriented security measures

§32.6 This brief analysis necessarily raises more questions than it can explore

§32.7 A familiar alternative scenario, global control by a value-aligned AGI agent, presents several difficulties

§33 Competitive AI capabilities will not be boxed

Thesis. Because the world’s aggregate AI capacity will greatly exceed that of any single system, the classic “AI confinement” challenge (with AI in a box and humans outside) is better regarded as an idealization than as a concrete problem situation.

§33.1 Summary

§33.2 SI-level capabilities will likely emerge from incremental R&D automation

§33.3 We can expect AI R&D capacity to be distributed widely, beyond any “box”

§33.4 AI systems will be instantiated together with diverse peer-level systems

§33.5 The ability to instantiate diverse, highly-capable systems presents both risks and opportunities for AI safety

§34 R&D automation is compatible with both strong and weak centralization

Thesis. Advances in AI R&D automation are currently distributed across many independent research groups, but a range of pressures could potentially lead to strong centralization of capabilities.

§34.1 Summary

§34.2 The R&D automation model is compatible with decentralized development

§34.3 Accelerating progress could lead to strong centralization of capabilities

§34.4 Centralization does not imply a qualitative change in R&D tasks

§34.5 Centralization and decentralization provide differing affordances relevant to AI policy and strategy

§35 Predictable aspects of future knowledge can inform AI safety strategies

Thesis. AI developers will accumulate extensive safety-relevant knowledge in the course of their work, and predictable aspects of that future knowledge can inform current studies of strategies for safe AI development.

§35.1 Summary

§35.2 Advanced AI systems will be preceded by similar but simpler systems

§35.3 Large-scale successes and failures rarely precede smaller successes and failures

§35.4 AI researchers eagerly explore and exploit surprising capabilities

§35.5 AI developers will be alert to patterns of unexpected failure

§35.6 AI safety researchers will be advising (responsible) AI developers

§35.7 Considerations involving future safety-relevant knowledge call for further exploration

§36 Desiderata and directions for interim AI safety guidelines

Thesis. Interim AI safety guidelines should (and could) engage with present practice, place little burden on practitioners, foster future safety-oriented development, and promote an ongoing process of guideline development and adoption.

§36.1 Summary

§36.2 Desiderata

§36.3 Good practice in development tends to align with safety concerns

§36.4 Exploring families of architectures and tasks builds practical knowledge

§36.5 Task-oriented development and testing improve both reliability and safety

§36.6 Modular architectures make systems more understandable and predictable

§36.7 Interim safety guidelines can foster ongoing progress

§37 How do neural and symbolic technologies mesh?

Thesis. Neural networks and symbolic/algorithmic AI technologies are complements, not alternatives; they are being integrated in multiple ways at levels that range from components and algorithms to system architectures.

§37.1 Summary

§37.2 Motivation

§37.3 A crisp taxonomy of NN and S/A systems is elusive and unnecessary

§37.4 NN and S/A techniques are complementary

§37.5 AI-service development can scale to comprehensive, SI-level services
§37.6 Integration at the level of components and mechanisms
§37.7 Integration at the level of algorithmic and representational structures
§37.8 Integration at the level of systems and subsystems
§37.9 Integration of NN and S/A techniques is a rich and active research frontier

§38 Broadly-capable systems coordinate narrower systems

Thesis. In both human and AI systems, we see broad competencies built on narrower competencies; this pattern of organization is a robust feature of intelligent systems, and scales to systems that deliver broad services at a superhuman level.

§38.1 Summary
§38.2 Today's superhuman competencies reside in organizational structures
§38.3 Specialization has robust advantages in learning diverse competencies
§38.4 Division of knowledge and labor is universal in performing complex tasks
§38.5 Current AI services show strong task differentiation
§38.6 AI systems trained on seemingly indivisible tasks learn to divide labor
§38.7 Black-box abstractions discard what we know about the architecture of systems with broad capabilities

§39 Tiling task-space with AI services can provide general AI capabilities

Thesis. Joint embedding of learned vector representations can be used to map tasks to services, enabling systems to provide general, extensible, seamlessly-integrated AI capabilities by exploiting the expertise of relatively narrow AI components.

§39.1 Summary
§39.2 Broad capabilities call for mechanisms that compose diverse competencies
§39.3 The task-space concept suggests a model of integrated AI services
§39.4 Embeddings in high-dimensional spaces provide powerful representations
§39.5 High-dimensional embeddings can represent semantically rich domains
§39.6 Proximity-based (application/activation/access) can deliver diverse services
§39.7 PBA operations are pervasive in deep learning systems
§39.8 Joint embeddings can link related semantic domains
§39.9 PBA operations can help match tasks to candidate services at scale
§39.10 PBA operations can help match new tasks to service-development services
§39.11 Integrated, extensible AI services constitute general artificial intelligence

§40 Could 1 PFLOP/s systems exceed the basic functional capacity of the human brain?

Thesis. Multiple comparisons between narrow AI tasks and narrow neural tasks concur in suggesting that PFLOP/s computational systems exceed the basic functional capacity of the human brain.

§40.1 Summary
§40.2 Metrics and methodology
§40.3 Estimated ratios for specific machine tasks
§40.4 Even with current methods, training can be fast by human standards
§40.5 Large computational costs for training need not substantially undercut the implications of low costs for applications
§40.6 Conclusions

Afterword

Acknowledgements

References