> FHI TECHNICAL REPORT <

Reframing Superintelligence

Comprehensive AI Services as General Intelligence

K. Eric Drexler

Technical Report #2019-1

Cite as:

Drexler, K.E. (2019): "Reframing Superintelligence: Comprehensive AI Services as General Intelligence", Technical Report #2019-1, Future of Humanity Institute, University of Oxford

Abstract

Studies of superintelligent-level systems have typically posited AI functionality that plays the role of a mind in a rational utility-directed agent, and hence employ an abstraction initially developed as an idealized model of human decision makers. Today, developments in AI technology highlight intelligent systems that are quite unlike minds, and provide a basis for a different approach to understanding them: Today, we can consider how AI systems are produced (through the work of research and development), what they do (broadly, provide services by performing tasks), and what they will enable (including incremental yet potentially thorough automation of human tasks).

Because tasks subject to automation include the tasks that comprise AI research and development, current trends in the field promise accelerating AI-enabled advances in AI technology itself, potentially leading to asymptotically recursive improvement of AI technologies in distributed systems, a prospect that contrasts sharply with the vision of self-improvement internal to opaque, unitary agents.

The trajectory of AI development thus points to the emergence of *asymptotically comprehensive, superintelligent-level AI services* that—crucially—can include the service of developing new services, both narrow and broad, guided by concrete human goals and informed by strong models of human (dis)approval. The concept of comprehensive AI services (CAIS) provides a model of flexible, general intelligence in which agents are a class of service-providing products, rather than a natural or necessary engine of progress in themselves.

Ramifications of the CAIS model reframe not only prospects for an intelligence explosion and the nature of advanced machine intelligence, but also the relationship between goals and intelligence, the problem of harnessing advanced AI to broad, challenging problems, and fundamental considerations in AI safety and strategy. Perhaps surprisingly, strongly self-modifying agents lose their instrumental value even as their implementation becomes more accessible, while the likely context for the emergence of such agents becomes a world already in possession of general superintelligent-level capabilities. These prospective capabilities, in turn, engender novel risks and opportunities of their own.

Further topics addressed in this work include the general architecture of systems with broad capabilities, the intersection between symbolic and neural systems, learning *vs.* competence in definitions of intelligence, tactical *vs.* strategic tasks in the context of human control, and estimates of the relative capacities of human brains *vs.* current digital systems.

Contents

Preface

The writing of this document was prompted by the growing gap between models that equate advanced AI with powerful agents and the emerging reality of advanced AI as an expanding set of capabilities (here, "services") in which agency is optional. A service-centered perspective reframes both prospects for superintelligent-level AI and a context for studies of AI safety and strategy.

Taken as a whole, this work suggests that problems centered on *what high-level AI systems might choose to do* are relatively tractable, while implicitly highlighting questions of *what humans might choose to do with their capabilities*. This shift, in turn, highlights the potentially pivotal role of high-level AI in solving problems created by high-level AI technologies themselves.

The text was written and shared as a series of widely-read Google Docs released between December 2016 and November 2018, largely in response to discussions within the AI safety community. The organization of the present document reflects this origin: The sections share a common conceptual framework, yet address diverse, overlapping, and often loosely-coupled topics. The table of contents, titles, subheads, summaries, and internal links are structured to facilitate skimming by readers with different interests. The table of contents primarily of declarative sentences, and has been edited to read as an overview.

Several apologies are in order: A number of topics and examples assume a basic familiarity with deep-learning concepts and jargon, while much of the content assumes familiarity with concerns regarding artificial general intelligence circa 2016–18; some sections directly address concepts and concerns discussed in *Superintelligence* (Bostrom 2014). In this work, I have made little effort to assign proper scholarly credit to ideas: Concepts that seem natural, obvious, or familiar are best treated as latent community knowledge and very likely have uncited antecedents. Ideas that can reasonably be attributed to someone else probably should be. Finally, how I frame and describe basic concepts has shifted over time, and in the interests of early completion, I have made only a modest effort to harmonize terminology across the original documents. I thought it best to share the content without months of further delay.

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

> 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

The emerging trajectory of AI development reframes AI prospects. Ongoing automation of AI R&D tasks, in conjunction with the expansion of AI services, suggests a tractable, non-agent-centric model of recursive AI technology improvement that can implement general intelligence in the form of comprehensive AI services (CAIS), a model that includes the service of developing new services. The CAIS model—which scales to superintelligent-level capabilities—follows software engineering practice in abstracting functionality from implementation while maintaining the familiar distinction between application systems and development processes. Language translation exemplifies a service that could incorporate broad, superintelligent-level world knowledge while avoiding classic AI-safety challenges both in development and in application. Broad world knowledge could likewise support predictive models of human concerns and (dis)approval, providing safe, potentially superintelligent-level mechanisms applicable to problems of AI alignment. Taken as a whole, the R&D-automation/CAIS model reframes prospects for the development and application of superintelligence, placing prospective AGI agents in the context of a broader range of intelligent systems while attenuating their marginal instrumental value.

I.2 The trajectory of AI development reframes AI prospects

Past, present, and projected developments in AI technology can inform our understanding of prospects for superintelligent-level capabilities, providing a concrete anchor that complements abstract models of potential AI systems. A development-oriented perspective highlights path-dependent considerations in assessing potential risks, risk-mitigation measures, and safety-oriented

research strategies. The current trajectory of AI development points to asymptotically recursive automation of AI R&D that can enable the emergence of general, asymptotically comprehensive AI services (CAIS). In the R&D-automation/CAIS model, recursive improvement and general AI capabilities need not be embodied in systems that act as AGI agents.

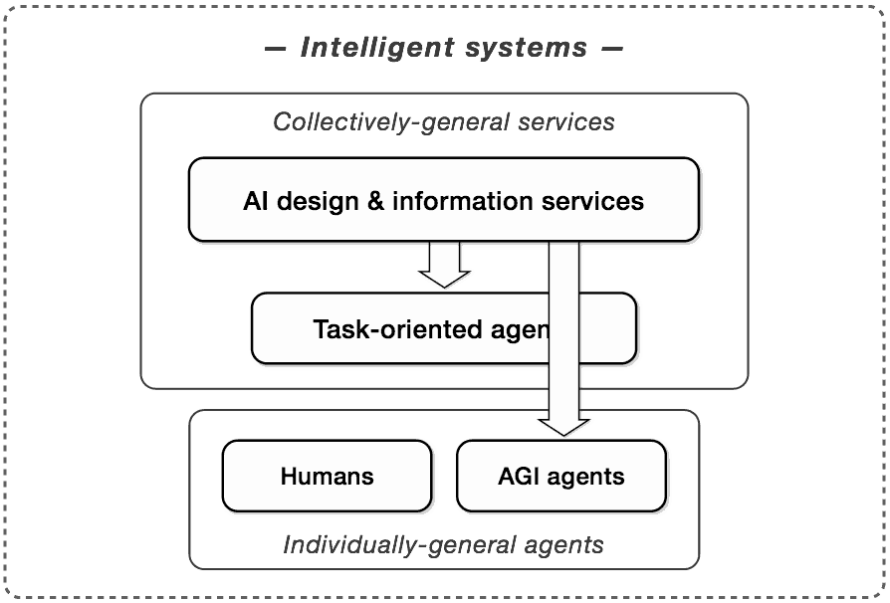

Figure 1: *Classes of intelligent systems*

I.3 R&D automation suggests a *technology-centered* model of recursive improvement

Technology improvement proceeds through research and development, a transparent process that exposes component tasks to inspection, refactoring, and incremental automation.[1] If we take advanced AI seriously, then accelerating, asymptotically-complete R&D automation is a natural consequence:

- By hypothesis, advances in AI will enable incremental automation and speedup of all human tasks.
- As-yet unautomated AI R&D tasks are human tasks, hence subject to incremental automation and speedup.
- Therefore, advances in AI will enable incremental automation and speedup of all AI R&D tasks.

1. Note that component-level opacity is compatible with effective system-level transparency.

Today we see automation and acceleration of an increasing range of AI R&D tasks, enabled by the application of both conventional software tools and technologies in the AI spectrum. Past and recent developments in the automation of deep-learning R&D tasks include:

- Diverse mechanisms embodied in NN toolkits and infrastructures
- Black-box and gradient-free optimization for NN hyperparameter search (Jaderberg et al. 2017)
- RL search and discovery of superior NN gradient-descent algorithms (Bello et al. 2017)
- RL search and discovery of superior NN cells and architectures (Zoph et al. 2017)

Today, automation of search and discovery (a field that overlaps with "meta-learning") requires human definition of search spaces, and we can expect that the *definition of new search spaces*—as well as fundamental innovations in architectures, optimization methods, and the definition and construction of tasks—will remain dependent on human insight for some time to come. Nonetheless, increasing automation of even relatively narrow search and discovery could greatly accelerate the implementation and testing of advances based on human insights, as well as their subsequent integration with other components of the AI technology base. Exploring roles for new components (including algorithms, loss functions, and training methods) can be routine, yet important: as Geoff Hinton has remarked, "A bunch of slightly new ideas that play well together can have a big impact".

Focusing exclusively on relatively distant prospects for *full* automation would distract attention from the potential impact of incremental research automation in accelerating automation itself.

I.4 R&D automation suggests a *service-centered* model of general intelligence

AI deployment today is dominated by AI services such as language translation, image recognition, speech recognition, internet search, and a host of services buried within other services. Indeed, corporations that provide cloud computing now actively promote the concept of "AI as a service" to other corporations. Even applications of AI within autonomous systems (*e.g.*, self-driving vehicles) can be regarded as providing services (planning, perception, guidance) to other system components.

R&D automation can itself be conceptualized as a set of services that directly or indirectly enable the implementation of new AI services. Viewing

service development through the lens of R&D automation, tasks for advanced AI include:

- Modeling human concerns
- Interpreting human requests
- Suggesting implementations
- Requesting clarifications
- Developing and testing systems
- Monitoring deployed systems
- Assessing feedback from users
- Upgrading and testing systems

CAIS functionality, which includes the service of developing stable, task-oriented AI agents, subsumes the instrumental functionality of proposed self-transforming AGI agents, and can present that functionality in a form that better fits the established conceptual frameworks of business innovation and software engineering.

I.5 The services model abstracts *functionality* from *implementation*

Describing AI systems in terms of functional behaviors ("services") aligns with concepts that have proved critical in software systems development. These include separation of concerns, functional abstraction, data abstraction, encapsulation, and modularity, including the use of client/server architectures—a set of mechanisms and design patterns that support effective program design, analysis, composition, reuse, and overall robustness.

Abstraction of functionality from implementation can be seen as a figure-ground reversal in systems analysis. Rather than considering a complex system and asking how it will behave, one considers a behavior and asks how it can be implemented. Desired behaviors can be described as services, and experience shows that complex services can be provided by combinations of more specialized service providers, some of which provide the service of aggregating and coordinating other service providers.

I.6 The R&D automation model distinguishes *development* from *functionality*

The AI-services model maintains the distinction between AI *development* and AI *functionality*. In the development-automation model of advanced AI ser-

vices, stable systems build stable systems, avoiding both the difficulties and potential dangers of building systems subject to open-ended self-transformation and potential instability.

Separating development from application has evident advantages. For one, task-focused applications need not themselves incorporate an AI-development apparatus—there is little reason to think that a system that provides online language translation or aerospace engineering design services should in addition be burdened with the tasks of an AI developer. Conversely, large resources of information, computation, and time can be dedicated to AI development, far beyond those required to perform a typical service. Likewise, in ongoing service application and upgrade, aggregating information from multiple deployed systems can provide decisive advantages to centralized development (for example, by enabling development systems for self-driving cars to learn from millions of miles of car-experience per day). Perhaps most important, stable products developed for specific purposes by a dedicated development process lend themselves to extensive pre-deployment testing and validation.

I.7 Language translation exemplifies a safe, potentially superintelligent service

Language translation provides an example of a service best provided by superintelligent-level systems with broad world knowledge. Translation of written language maps input text to output text, a bounded, episodic, sequence-to-sequence task. Training on indefinitely large and broad text corpora could improve translation quality, as could deep knowledge of psychology, philosophy, history, geophysics, chemistry, and engineering. Effective optimization of a translation system for an objective that weights both quality and efficiency would focus computation solely on the application of this knowledge to translation.

The process of *developing* language translation systems is itself a service that can be formulated as an episodic task, and as with translation itself, effective optimization of translation-development systems for both quality and efficiency would focus computation solely on that task.

There is little to be gained by modeling stable, episodic service-providers as rational agents that optimize a utility function over future states of the world, hence a range of concerns involving utility maximization (to say nothing of self-transformation) can be avoided across a range of tasks. Even superintelligent-level world knowledge and modeling capacity need not in itself lead to strategic behavior.

I.8 Predictive models of human (dis)approval can aid AI goal alignment

As noted by Stuart Russell, written (and other) corpora provide a rich source of information about human opinions regarding actions and their effects; intelligent systems could apply this information in developing predictive models of human approval, disapproval, and disagreement. Potential training resources for models of human approval include existing corpora of text and video, which reflect millions of person-years of both real and imagined actions, events, and human responses; these corpora include news, history, fiction, science fiction, advice columns, law, philosophy, and more, and could be augmented and updated with the results of crowd-sourced challenges structured to probe model boundaries.

Predictive models of human evaluations could provide strong priors and common-sense constraints to guide both the implementation and actions of AI services, including strategic advisory services to powerful actors. Predictive models are not themselves rational agents, yet models of this kind could contribute to the solution of agent-centered safety concerns. In this connection, separation of development from application can insulate such models from perverse feedback loops involving self-modification.

I.9 The R&D-automation/CAIS model reframes prospects for superintelligence

From a broad perspective, the R&D-automation/CAIS model:

- Distinguishes recursive technology improvement from self-improving agents
- Shows how incremental automation of AI R&D can yield recursive improvement
- Presents a model of general intelligence centered on services rather than systems
- Suggests that AGI agents are not necessary to achieve instrumental goals
- Suggests that high-level AI services would precede potential AGI agents
- Suggests potential applications of high-level AI services to general AI safety

For the near term, the R&D-automation/CAIS model:

- Highlights opportunities for safety-oriented differential technology development

- Highlights AI R&D automation as a leading indicator of technology acceleration
- Suggests rebalancing AI research portfolios toward AI-enabled R&D automation

Today, we see strong trends toward greater AI R&D automation and broader AI services. We can expect these trends to continue, potentially bridging the gap between current and superintelligent-level AI capabilities. Realistic, path-dependent scenarios for the emergence of superintelligent-level AI capabilities should treat these trends both as an anchor for projections and as a prospective context for trend-breaking developments.

II Overview: Questions, propositions, and topics

II.1 Summary

This document outlines topics, questions, and propositions that address:

1. Prospects for an intelligence explosion
2. The nature of advanced machine intelligence
3. The relationship between goals and intelligence
4. The problem of using and controlling advanced AI
5. Near- and long-term considerations in AI safety and strategy

The questions and propositions below reference sections of this document that explore key topics in more depth. From the perspective of AI safety concerns, this document offers support for several currently-controversial propositions regarding artificial general intelligence:

- That AGI agents have no natural role in developing general AI capabilities.
- That AGI agents would offer no unique and substantial value in providing general AI services.
- That AI-based security services could safely constrain subsequent AGI agents, even if these operate at a superintelligent level.

II.2 Reframing prospects for an intelligence explosion

II.2.1 Does recursive improvement imply self-transforming agents?

Ongoing automation of tasks in AI R&D suggests a model of asymptotically-recursive technology improvement that scales to superintelligent-level (SI-level) systems. In the R&D-automation model, recursive improvement is systemic, not internal to distinct systems or agents. The model is fully generic: It requires neither assumptions regarding the content of AI technologies, nor assumptions regarding the pace or sequence of automation of specific R&D tasks. Classic self-transforming AGI agents would be strictly more difficult to implement, hence are not on the short path to an intelligence explosion.

- *Section 1: R&D automation provides the most direct path to an intelligence explosion*

- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*

II.2.2 Would self-transforming agents provide uniquely valuable functionality?

Suites of AI services that support SI-level AI development—working in consultation with clients and users—could provide a comprehensive range of novel AI services; these would presumably include services provided by adaptive, upgradable, task-oriented agents. It is difficult to see how the introduction of potentially unstable agents that undergo autonomous open-ended self-transformation would provide additional value.

- *Section 3: To understand AI prospects, focus on services, not implementations*
- *Section 12: AGI agents offer no compelling value*

II.2.3 Can fast recursive improvement be controlled and managed?

Recursive improvement of *basic AI technologies* would apply allocated machine resources to the development of increasingly functional building blocks for AI applications (better algorithms, architectures, training methods, *etc.*); basic technology development of this sort could be open-ended, recursive, and fast, yet non-problematic. *Deployed AI applications* call for careful management, but applications stand outside the inner loop of basic-technology improvement.

- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*

II.2.4 Would general learning algorithms produce systems with general competence?

The application of an idealized, fully general learning algorithm would *enable* but not *entail* the learning of any particular competence. Time, information, and resource constraints are incompatible with universal competence, regardless of *ab initio* learning capacity.

- *Section 2: Standard definitions of "superintelligence" conflate learning with competence*
- *Section 21: Broad world knowledge can support safe task performance*

- *Section 39: Tiling task-space with AI services can provide general AI capabilities*

II.3 Reframing the nature of advanced machine intelligence

II.3.1 Is human learning an appropriate model for AI development?

Action, experience, and learning are typically decoupled in AI development: Action and experience are aggregated, not tied to distinct individuals, and the machine analogues of cognitive change can be profound during system development, yet absent in applications. As we see in AI technology today, learning algorithms can be applied to produce and upgrade systems that do not themselves embody those algorithms. Accordingly, using human learning and action as a model for AI development and application can be profoundly misleading.

- *Section 2: Standard definitions of "superintelligence" conflate learning with competence*
- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 16: Aggregated experience and centralized learning support AI-agent applications*

II.3.2 Does stronger optimization imply greater capability?

Because optimization for a task focuses capabilities on that task, strong optimization of a system acts as a strong constraint; in general, optimization does not extend the scope of a task or increase the resources employed to perform it. System optimization typically tends to reduce resource consumption, increase throughput, and improve the quality of results.

- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*

II.3.3 Do broad knowledge and deep world models imply broad AI capabilities?

Language translation systems show that safe, stable, high-quality task performance can be compatible with (and even require) broad and deep knowledge about the world. The underlying principle generalizes to a wide range of tasks.

- *Section 21: Broad world knowledge can support safe task performance*

II.3.4 Must we model SI-level systems as rational, utility-maximizing agents?

The concept of rational, utility-maximizing agents was developed as an idealized model of human decision makers, and hence is inherently (though abstractly) anthropomorphic. Utility-maximizing agents may be intelligent systems, but intelligent systems (and in particular, *systems of agents*) need not be utility-maximizing agents.

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 6: A system of AI services is not equivalent to a utility maximizing agent*
- *Section 17: End-to-end reinforcement learning is compatible with the AI-services model*
- *Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents*

II.3.5 Must we model SI-level systems as unitary and opaque?

Externally-determined features of AI components (including their development histories, computational resources, communication channels, and degree of mutability) can enable structured design and functional transparency, even if the components themselves employ opaque algorithms and representations.

- *Section 9: Opaque algorithms are compatible with functional transparency and control*
- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 38: Broadly-capable systems coordinate narrower systems*

II.4 Reframing the relationship between goals and intelligence

II.4.1 What does the orthogonality thesis imply for the generality of convergent instrumental goals?

Intelligent systems optimized to perform perform bounded tasks (in particular, episodic tasks with a bounded time horizon) need not be agents with open-ended goals that call for self preservation, cognitive enhancement, resource acquisition, and so on; by Bostrom's orthogonality thesis, this holds true regardless of the level of intelligence applied to those tasks.

- *Section 19: The orthogonality thesis undercuts the generality of instrumental convergence*

II.4.2 How broad is the basin of attraction for convergent instrumental goals?

Instrumental goals are closely linked to final goals of indefinite scope that concern the indefinite future. Societies, organizations, and (in some applications) high-level AI agents may be drawn toward convergent instrumental goals, but high-level intelligence *per se* does not place AI systems within this basin of attraction, even if applied to broad problems that are themselves long-term.

- *Section 21: Broad world knowledge can support safe task performance*
- *Section 25: Optimized advice need not be optimized to induce its acceptance*

II.5 Reframing the problem of using and controlling advanced AI

II.5.1 Would the ability to implement potentially-risky self-transforming agents strongly motivate their development?

If future AI technologies could implement potentially-risky, self-transforming AGI agents, then similar, more accessible technologies could more easily be applied to implement open, comprehensive AI services. Because the service of providing new services subsumes the proposed instrumental value of self-transforming agents, the incentives to implement potentially-risky self-transforming agents appear to be remarkably small.

- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*
- *Section 12: AGI agents offer no compelling value*
- *Section 33: Competitive AI capabilities will not be boxed*

II.5.2 How can human learning during AI development contribute to current studies of AI safety strategies?

We can safely predict that AI researchers will continue to identify and study surprising AI behaviors, and will seek to exploit, mitigate, or avoid them in developing AI applications. This and other predictable aspects of future knowledge can inform current studies of strategies for safe AI development.

- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*

II.5.3 Can we avoid strong trade-offs between development speed and human oversight?

Basic research, which sets the overall pace of technological progress, could be safe and effective with relatively little human guidance; application development, by contrast, requires strong human guidance, but as an inherent part of the development task—to deliver desirable functionality—rather than as an impediment. Support for human guidance can be seen as an AI service, and can draw on predictive models of human approval and concerns.

- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*

II.5.4 Can we architect safe, superintelligent-level design and planning services?

Consideration of concrete task structures and corresponding services suggests that SI-level AI systems can safely converse with humans, perform creative search, and propose designs for systems to be implemented and deployed in the world. Systems that provide design and planning services can be optimized to provide advice without optimizing to manipulate human acceptance of that advice.

- *Section 26: Superintelligent-level systems can safely provide design and planning services*
- *Section 28: Automating biomedical R&D does not require defining human welfare*

II.5.5 Will competitive pressures force decision-makers to transfer strategic decisions to AI systems?

In both markets and battlefields, advantages in reaction time and decision quality can motivate transfer of tactical control to AI systems, despite potential risks; for strategic decisions, however, the stakes are higher, speed is less important, advice can be evaluated by human beings, and the incentives to yield control are correspondingly weak.

- *Section 27: Competitive pressures provide little incentive to transfer strategic decisions to AI systems*

II.5.6 What does the R&D-automation/AI-services model imply for studies of conventional vs. extreme AI-risk concerns?

Increasing automation of AI R&D suggests that AI capabilities may advance surprisingly rapidly, a prospect that increases the urgency of addressing conventional AI risks such as unpredictable failures, adversarial manipulation, criminal use, destabilizing military applications, and economic disruption. Prospects for the relatively rapid emergence of systems with broad portfolios of capabilities, including potentially autonomous planning and action, lend increased credence to extreme AI-risk scenarios, while the AI-services model suggests strategies for avoiding or containing those risks while gaining the benefits of high- and SI-level AI capabilities.

- *Section 12: AGI agents offer no compelling value*
- *Section 14: The AI-services model brings ample risks*

II.5.7 What can agent-oriented studies of AI safety contribute, if risky AI agents are optional?

People will want AI systems that plan and act in the world, and some systems of this class can naturally be modeled as rational, utility-directed agents. Studies of systems within the rational-agent model can contribute to AI safety and strategy in multiple ways, including:

- Expanding the range of safe AI-agent architectures by better understanding how to define bounded tasks in a utility-directed framework.
- Expanding safe applications of utility-directed agents to less well-bounded tasks by better understanding how to align utility functions with human values.
- Better understanding the boundaries beyond which combinations of agent architectures and tasks could give rise to classic AGI-agent risks.
- Better understanding how (and under what conditions) evolutionary pressures could engender perverse strategic behavior in nominally non-agent-like systems.
- Exploring ways access to high-level AI services could could help to avoid or mitigate classic agent-centered AI risks.

- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 29: The AI-services model reframes the potential* roles *of AGI agents*

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

II.6.1 Could widely available current or near-term hardware support superintelligence?

Questions of AI safety and strategy become more urgent if future, qualitatively SI-level computation can be implemented with greater-than-human task throughput on affordable, widely-available hardware. There is substantial reason to think that this condition already holds.

- *Section 40: Could 1 PFLOP/s systems exceed the basic functional capacity of the human brain?*

II.6.2 What kinds of near-term safety-oriented guidelines might be feasible and useful?

Current technology presents no catastrophic risks, and several aspects of current development practice align not only with safety, but with good practice in science and engineering. Development of guidelines that codify current good practice could contribute to near-term AI safety with little organizational cost, while also engaging the research community in an ongoing process that addresses longer-term concerns.

- *Section 4: The AI-services model includes both descriptive and prescriptive aspects*
- *Section 34: R&D automation is compatible with both strong and weak centralization*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*
- *Section 36: Desiderata and directions for interim AI safety guidelines*

II.6.3 How can near-term differential technology development contribute to safety?

Concerns with AI safety and strategy should influence research agendas intended to promote broad societal benefits. Directions that deserve emphasis include work on concrete problems in AI safety, on predictive models of human approval and disapproval, and on capabilities that could facilitate the detection of potential bad actors.

- *Section 22: Machine learning can develop predictive models of human approval*

- *Section 23: AI development systems can support effective human guidance*

II.6.4 In the long term, are unaligned superintelligent agents compatible with safety?

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

- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 30: Risky AI can help develop safe AI*
- *Section 31: Supercapabilities do not entail "superpowers"*
- *Section 32: Unaligned superintelligent agents need not threaten world stability*

II.7 Conclusions

We can expect to see AI-enabled automation of AI research and development continue to accelerate, both leveraging and narrowing the scope of human insights required for progress in AI technologies. Asymptotically-recursive improvement of AI technologies can scale to a superintelligent level, supporting the development of a fully-general range of high-level AI services that includes the service of developing new services in response to human demand. Because general AI-development capabilities do not imply general capabilities in any particular system or agent, classic AGI agents would be potential *products* of SI-level AI development capabilities, not a path to uniquely valuable functionality.

Within the space of potential intelligent systems, agent-centered models span only a small region, and even abstract, utility-directed rational-agent models invite implicitly anthropomorphic assumptions. In particular, taking human learning as a model for machine learning has encouraged the conflation of intelligence-as-learning-capacity with intelligence-as-competence, while these aspects of intelligence are routinely and cleanly separated AI system development: Learning algorithms are typically applied to train systems that do not themselves embody those algorithms.

The service-centered perspective on AI highlights the generality of Bostrom's Orthogonality Thesis: SI-level capabilities can indeed be applied to any task, including services that are (as is typical of services) optimized to deliver bounded results with bounded resources in bounded times. The pursuit of Bostrom's convergent instrumental goals would impede—not

improve—the performance of such services, yet would be natural for those same instrumental goals to be hotly pursued by human organizations (or AI agents) that *employ* AI services.

Prospects for service-oriented superintelligence reframe the problem of managing advanced AI technologies: Potentially-risky self-transforming agents become optional, rather than overwhelmingly valuable, the separation of basic research from application development can circumvent trade-offs between development speed and human oversight, and natural task architectures suggest safe implementations of SI-level design and planning services. In this connection, distinctions between tactical execution and strategic advice suggest that even stringent competitive pressures need not push decision-makers to cede strategic decisions to AI systems.

In contrast to unprecedented-breakthrough models that postulate runaway self-transforming agents, prospects for the incremental emergence of diverse, high-level AI capabilities promise broad, safety-relevant experience with problematic (yet not catastrophic) AI behaviors. Safety guidelines can begin by codifying current safe practices, which include training and re-training diverse architectures while observing and studying surprising behaviors. The development of diverse, high-level AI services also offers opportunities for safety-relevant differential technology development, including the development of common-sense predictive models of human concerns that can be applied to improve the value and safety of AI services and AI agents.

The R&D-automation/AI-services model suggests that conventional AI risks (*e.g.*, failures, abuse, and economic disruption) are apt to arrive more swiftly than expected, and perhaps in more acute forms. While this model suggests that extreme AI risks may be relatively avoidable, it also emphasizes that such risks could arise more quickly than expected. In this context, agent-oriented studies of AI safety can both expand the scope of safe agent applications and improve our understanding of the conditions for risk. Meanwhile, service-oriented studies of AI safety could usefully explore potential applications of high-level services to general problems of value alignment and behavioral constraint, including the potential architecture of security services that could ensure safety in a world in which some extremely intelligent agents are not inherently trustworthy.

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

> 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

AI-enabled development of improved AI algorithms could potentially lead to an accelerating feedback process, enabling a so-called "intelligence explosion". This quite general and plausible concept has commonly been identified with a specific, challenging, and risky implementation in which a predominant concentration of AI-development functionality is embodied in a distinct, goal-directed, self-transforming AI system—an AGI agent. A task-oriented analysis of AI-enabled AI development, however, suggests that self-transforming agents would play no natural role, even in the limiting case of explosively-fast progress in AI technology. Because the mechanisms underlying a potential intelligence explosion are already in operation, and have no necessary connection to unprecedented AGI agents, paths to extremely rapid progress in AI technology may be both more direct and more controllable than has been commonly assumed.

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

As suggested by Good (1966), AI systems could potentially outperform human beings in the task of AI development, and hence "could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind."

1.3 Risk models have envisioned AGI agents driving an intelligence explosion

The widespread (yet questionable) assumption that this feedback process—"recursive improvement"—would entail self-transformation of a particular AI

system that engages with the world as an agent[1] has motivated a threat model in which "the machine" might not be "docile enough to tell us how to keep it under control" (Good 1966, p.33). The surprisingly complex and difficult ramifications of this threat model have been extensively explored in recent years (*e.g.*, in Bostrom 2014).

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

Advances in AI technology emerge from research and development,[2] a process that comprises a range of different technical tasks. These tasks are loosely coupled, and none requires universal competence: Consider, for example, the technology-centered tasks of training-algorithm development, benchmark development, architecture search, and chip design; and beyond these, application development tasks that include human consultation[3] and application prototyping, together with testing, monitoring, and customer service.[4]

Accordingly, general-purpose, self-transforming AI agents play no natural role in the process of AI R&D: They are potential (and potentially dangerous) *products*—not *components*—of AI development systems. To the extent that the concept of "docility" may be relevant to development systems as a whole, this desirable property is also, by default, deeply ingrained in the nature of the services they provide.[5]

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

It may be tempting to imagine that *self*-improvement would be simpler than loosely-coupled *systemic* improvement, but drawing a conceptual boundary around a system does not simplify its contents, and to require that systems capable of open-ended AI development *also* exhibit tight integration, functional autonomy, and operational agency would increase—not reduce—

1. Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame
2. Section 10: R&D automation dissociates recursive improvement from AI agency
3. Section 23: AI development systems can support effective human guidance
4. Section 16: Aggregated experience and centralized learning support AI-agent applications
5. Section 23: AI development systems can support effective human guidance

implementation challenges.[1] To attempt to place competitive R&D capabilities inside an agent presents difficulties, yet provides no compensating advantages in performance or utility.[2]

The pervasive assumption that an intelligence explosion must await the development of agents capable of autonomous, open-ended self improvement has encouraged skepticism and complacency[3] regarding prospects for superintelligent-level AI.[4] If we think that any given set of human tasks can be automated, however, than so can any—and in the limit, all—AI R&D tasks. This proposition seems relatively uncontroversial, yet has profound implications for the likely trajectory of AI development.

AI-enabled automation of fundamental AI R&D tasks is markedly accelerating.[5] As the range of automated tasks increases, we can expect feedback loops to tighten, enabling AI development at a pace that, while readily controlled, has no obvious limits.

Further Reading

- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*
- *Section 12: AGI agents offer no compelling value*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 30: Risky AI can help develop safe AI*
- *Section 40: Could 1 PFLOP/s systems exceed the basic functional capacity of the human brain?*

1. Section 11: Potential AGI-enabling technologies also enable comprehensive AI services
2. Section 12: AGI agents offer no compelling value
3. See Chollet 2017
4. See Kelly 2017
5. For example, see Zoph et al. 2017; Bello et al. 2017; Jaderberg et al. 2017; Chen et al. 2017; Schrimpf et al. 2017

2 Standard definitions of "superintelligence" conflate learning with competence

> 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

Since Good (1966), superhuman intelligence has been equated with superhuman intellectual competence, yet this definition misses what we mean by *human* intelligence. A child is considered intelligent because of learning capacity, not competence, while an expert is considered intelligent because of competence, not learning capacity. Learning capacity and competent performance are distinct characteristics in human beings, and are routinely separated in AI development. Distinguishing *learning* from *competence* is crucial to understanding both prospects for AI development and potential mechanisms for controlling superintelligent-level AI systems.

2.2 Superintelligence has been defined in terms of adult human competence

Good (1966) defined "ultraintelligence" in terms of distinct, highly competent entities:

> *Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever.*

The standard definition of "superintelligence" today (Bostrom 2014) parallels Good (1966), yet to define *superintelligence* in terms of adult intellectual competence fails to capture what we mean by *human* intelligence.

2.3 "Intelligence" often refers instead to learning capacity

Consider what we mean when we call a person intelligent:

- A child is considered "intelligent" because of learning capacity, not competence.

- An expert is considered "intelligent" because of competence, not learning capacity.

We can overlook this distinction in the human world because learning and competence are deeply intertwined in human intellectual activities; in considering prospects for AI, by contrast, regarding "intelligence" as entailing both learning and competence invites deep misconceptions.

2.4 Learning and competence are separable in principle and practice

A human expert in science, engineering, or AI research might provide brilliant solutions to problems, and even if a drug or neurological defect blocked the expert's formation of long-term memories, we would recognize the expert's intelligence. Thus, even in humans, competence can in principle be dissociated from ongoing learning, while in AI technology, this separation is simply standard practice. Regardless of implementation technologies, each released version of an AI system can be a fixed, stable software object.

Reinforcement learning agents illustrate the separation between learning and competence: Reinforcement "rewards" are signals that shape learned behavior, yet play no role in performance. Trained RL agents exercise their competencies without receiving reward.[1]

2.5 Patterns of AI learning and competence differ radically from humans'

AI systems and human beings differ radically in how learning, knowledge transfer, competence, and experience are connected:

Aspect	In humans	In AI systems
Performance and learning:	Unified	Separable
Nature of experience:	Individual, sequential	Aggregated, parallel
Learning from experience:	Inherent	Optional
Knowledge transfer:	Instruction, study	Download

In short, AI systems can act without learning, learn without acting, and transfer knowledge without instruction or study; further, machine learning

1. Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents

can draw on the aggregated, parallel experience[1] of indefinitely large numbers of performing systems, which can then be upgraded by downloading new software.

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

Ensuring relatively predictable, constrained behavior is fundamental to AI control. The tacit assumption that the exercise of competence entails learning implies that an intelligent system must be mutable, which is to say, potentially unstable. Further, the tacit assumption that intelligence entails both learning and competence invites the misconception that AI systems capable of learning will necessarily have complex states and capabilities that might be poorly understood.

As this might suggest, conflating intelligence, competence, and learning obscures much of the potential solution space for of AI control problems (*e.g.*, approaches to architecting trustworthy composite oracles).[2] This problematic, subtly anthropomorphic model of intelligence is deeply embedded in current discussion of AI prospects. If we are to think clearly about AI control problems, then even comprehensive, superintelligent-level AI capabilities[3] must not be equated with "superintelligence" as usually envisioned.

Further Reading

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 16: Aggregated experience and centralized learning support AI-agent applications*
- *Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 30: Risky AI can help develop safe AI*

1. Section 16: Aggregated experience and centralized learning support AI-agent applications
2. Section 20: Collusion among superintelligent oracles can readily be avoided
3. Section 12: AGI agents offer no compelling value

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

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

From an instrumental perspective, intelligence centers on capabilities rather than systems, yet models of advanced AI commonly treat the key capability of *general intelligence*—the ability to develop novel capabilities—as a black-box mechanism embedded in a particular kind of system, an AGI agent. Service-oriented models of general intelligence instead highlight the *service of developing new services* as a computational action, and place differentiated task functionality (rather than unitary, general-purpose systems) at the center of analysis, linking models to the richly-developed conceptual framework of software engineering. Service-oriented models reframe the problem of aligning AI functionality with human goals, providing affordances absent from opaque-agent models.

3.2 The instrumental function of AI technologies is to provide services

In the present context, "services" are tasks performed to serve a client. AI systems may provide services to humans more-or-less directly (driving, designing, planning, conversing. . .), but in a software engineering context, one may also refer to clients and service-providers (servers) when both are computational processes.

Not every action performs a service. As with human intelligence, intelligence embodied in autonomous systems might not serve the instrumental goals of any external client; in considering actions of the system itself, the concept of goals and utility functions would then be more appropriate than the concept of services.

3.3 General intelligence is equivalent to general capability development

General intelligence requires an open-ended ability to develop novel capabilities, and hence to perform novel tasks. For instrumental AI, *capabilities* correspond to *potential services*, and general artificial intelligence can be regarded as an open-ended service able to provide new AI services by developing new capabilities.

3.4 The *ability to learn* is a capability

In humans (with their opaque skulls and brains), it is natural to distinguish internal capabilities (*e.g.*, learning to program) from external capabilities (*e.g.*, programming a machine), and to treat these as different in kind. In computational systems, however, there need be no such distinction: To develop a capability is to implement a system that provides that capability, a process that need not modify the system that performs the implementation. Removing artificial distinctions between kinds of capabilities improves the scope, generality, and homogeneity of models of artificial general intelligence.

3.5 Implementing new capabilities does not require "self modification"

An incremental computational process may extend the capabilities of a computational system. If the resulting code automatically replaces the previous version, and if it is convenient to regard that process as internal to the system, then it may be natural to call the process "self modification". In many practical applications, however, a different approach will produce better results: Data can be aggregated from many instances of a system, then combined through a centralized, perhaps computationally-intensive development service to produce upgraded systems that are then tested and deployed. The strong advantages of data aggregation and centralized development[1] suggest that it would be a mistake to adopt "self modification" as a default model of system improvement.

1. Section 16: Aggregated experience and centralized learning support AI-agent applications.

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

Services have structure: Across both human and computational worlds, we find that high-level services are provided by employing and coordinating lower-level services. We can expect that services of different kinds (*e.g.*, language translation, theorem proving, aircraft design, computer hacking, computer security, military strategy) will or readily could be developed and implemented as substantially distinct computational systems, each operating not only as a server, but as a client that itself employs a range of narrower services. It is safe to assert that the architecture of complex services and service-providers will be neither atomized nor monolithic.

3.7 Service-centered models harmonize with practice in software engineering

AI services are being developed and deployed in the context of other software services. Decades of research, billions of dollars, and enormous intellectual effort have been invested in organizing the development of increasingly complex systems, and universal patterns have emerged; in particular, system architectures are defined by their functions and interfaces—by service provided and means of employing them. The art of decomposing high-level functions into lower-level functions has been essential to making systems comprehensible, implementable, and maintainable. Modern software services are both the technological milieu of modern AI and a model for how complex information services emerge and evolve.

Perhaps the most fundamental principles are modularity and abstraction: partitioning functionality and then decoupling functionality from implementation. We can expect that these extraordinarily general abstractions can and will scale to systems implemented by (and to provide) superintelligent-level services.

3.8 Service-centered AI architectures can facilitate AI alignment

A system-centric model would suggest that general-purpose artificial intelligence must be a property of general-purpose AI systems, and that a fully-general AI system, to perform its functions, must be a powerful superintelligent agent. From this model and its conclusion, profound challenges follow.

A service-centric model, by contrast, proposes to satisfy general demands for intelligent services through a general capacity to develop services. This

general capacity is not itself a thing or an agent, but a pool of functionality that can be provided by coordination of AI-development services. In this model, even highly-capable agents implemented at a superintelligent level[1] can be stable, and need not themselves embody AI-development functionality. This model suggests that a range of profound challenges, if recognized, can also be avoided.

Diverse, potentially superintelligent-level AI services could be coordinated to provide the service of developing new AI services. Potential components and functions include:

- Predictive models of human approval, disapproval, and controversies.[2]
- Consulting services that propose and discuss potential products and services.[3]
- Design,[4] implementation,[5] and optimization services.[6]
- Specialists in technical security and safety measures.[7].
- Evaluation through criticism and red-team/blue-team competitions.[8]
- Pre-deployment testing and post-deployment assessment.[9]
- Iterative, experience-based upgrades to products and services.[10]

Each of the above corresponds to one or more high-level services that would typically rely on others, whether these are narrower (*e.g.*, language understanding and technical domain knowledge) or at a comparable level (*e.g.*, predictive models of human (dis)approval). Some services (*e.g.*, criticism and red-team/blue-team competitions) by nature interact with others that are adversarial and operationally distinct. Taken together, these services suggest

1. Section 29: The AI-services model reframes the potential *roles* of AGI agents.
2. Section 22: Machine learning can develop predictive models of human approval.
3. Section 23: AI development systems can support effective human guidance.
4. Section 26: Superintelligent-level systems can safely provide design and planning services.
5. Section 10: R&D automation dissociates recursive improvement from AI agency.
6. Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects.
7. Section 26: Superintelligent-level systems can safely provide design and planning services.
8. Section 20: Collusion among superintelligent oracles can readily be avoided.
9. Section 16: Aggregated experience and centralized learning support AI-agent applications.
10. Section 16: Aggregated experience and centralized learning support AI-agent applications.

a range of alignment-relevant affordances that are (to say the least) not salient in models that treat general intelligence as a black-box mechanism that is embedded in a general-purpose agent.

Further Reading

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*
- *Section 12: AGI agents offer no compelling value*
- *Section 16: Aggregated experience and centralized learning support AI-agent applications*
- *Section 29: The AI-services model reframes the potential* roles *of AGI agents*
- *Section 38: Broadly-capable systems coordinate narrower systems*

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

> 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

Does the AI-services model *describe prospects* for high-level AI application development, or *prescribe strategies* for avoiding classic AGI-agent risks? Description and prescription are closely aligned: The services model *describes* current and emerging patterns of AI application development, notes these patterns are accessible, scalable, and align with AI safety, and accordingly *prescribes* deliberate adherence to these patterns. The alignment between descriptive and prescriptive aspects of the services model is fortunate, because strategies for AI safety will be more readily adopted if they align with, rather than impede, the momentum of AI development.

4.2 The AI-services model describes current AI development

AI technology today advances through increasingly automated AI research and development[1], and produces applications that provide services[2], performing tasks such as translating languages, steering cars, recognizing faces, and beating Go masters. AI development itself employs a growing range of AI services, including architecture search, hyperparameter search, and training set development.

4.3 AI-service development scales to comprehensive, SI-level services

The "AI services" concept scales to sets of services that perform an asymptotically-comprehensive range of tasks, while AI-supported automation of AI R&D automation scales to asymptotically-recursive, potentially swift technology improvement. Because systems based on AI services (including service-development services) scale to a superintelligent level[3], the potential scope of AI services subsumes the instrumental functionality[4] that might otherwise motivate the development of AGI agents.[5]

4.4 Adherence to the AI-services model aligns with AI safety

Because the AI-services model naturally employs diversity, competition, and adversarial goals[6] (*e.g.*, proposers *vs.* critics) among service-providers, architectures that adhere to the (extraordinarily flexible) AI-services model can readily avoid classic risks associated with superintelligent, self-modifying, utility-maximizing agents.[7]

4.5 Adherence to the AI-services model seems desirable, natural,

1. Section 10: R&D automation dissociates recursive improvement from AI agency.
2. Section 3: To understand AI prospects, focus on services, not implementations.
3. Section 1: R&D automation provides the most direct path to an intelligence explosion.
4. Section 12: AGI agents offer no compelling value.
5. Section 11: Potential AGI-enabling technologies also enable comprehensive AI services.
6. Section 20: Collusion among superintelligent oracles can readily be avoided.
7. Section 6: A system of AI services is not equivalent to a utility maximizing agent

and practical

There need be no technological discontinuities on the way to thorough AI R&D automation and comprehensive AI services, and continued adherence to this model is compatible with efficient development and application of AI capabilities.[1] Traditional models of general AI capabilities, centered on AGI agents, seem more difficult to implement, more risky, and no more valuable in application. Accordingly, guidelines that prescribe adherence to the AI-services model[2] could improve[3] prospects for a safe path to superintelligent-level AI without seeking to impede the momentum of competitive AI development.

Further Reading

- *Section 1: R&D automation provides the most direct path to an intelligence explosion*
- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*
- *Section 12: AGI agents offer no compelling value*
- *Section 14: The AI-services model brings ample risks*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 36: Desiderata and directions for interim AI safety guidelines*

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

> 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

Concepts of artificial intelligence have long been tied to concepts of mind, and even abstract, rational-agent models of intelligent systems are built on

1. Section 24: Human oversight need not impede fast, recursive AI technology improvement.
2. Section 24: Human oversight need not impede fast, recursive AI technology improvement.
3. Section 14: The AI-services model brings ample risks.

psychomorphic and recognizably anthropomorphic foundations. Emerging AI technologies do not fit a psychomorphic frame, and are radically unlike evolved intelligent systems, yet technical analysis of prospective AI systems has routinely adopted assumptions with recognizably biological characteristics. To understand prospects for AI applications and safety, we must consider not only psychomorphic and rational-agent models, but also a wide range of intelligent systems that present strongly contrasting characteristics.

5.2 The concept of mind has framed our concept of intelligence

Minds evolved to guide organisms through life, and natural intelligence evolved to make minds more effective. Because the only high-level intelligence we know is an aspect of human minds, it is natural for our concept of mind to frame our concept of intelligence. Indeed, popular culture has envisioned advanced AI systems as artificial minds that are by default much like our own. AI-as-mind has powerful intuitive appeal.

The concept of AI-as-mind is deeply embedded in current discourse. For example, in cautioning against anthropomorphizing superintelligent AI, Bostrom (2014, p.105) urges us to "reflect for a moment on the vastness of the space of possible minds", an abstract space in which "human minds form a tiny cluster". To understand prospects for superintelligence, however, we must consider a broader space of potential intelligent systems, a space in which *mind-like systems themselves* form a tiny cluster.

5.3 Studies of advanced AI often posit intelligence in a psychomorphic role

Technical studies of AI control set aside explicitly human psychological concepts by modeling AI systems as goal-seeking rational agents. Despite their profound abstraction, however, rational-agent models originated as idealizations of human decision-making, and hence place intelligence in an implicitly anthropomorphic frame. More concretely, in rational-agent models, the *content* of human minds (human values, goals, cognitive limits…) is abstracted away, yet the *role* of minds in guiding decisions is retained. An agent's decision-making process fills an inherently mind-shaped slot, and that slot frames a recognizably psychomorphic concept of intelligence.

This is problematic: Although the rational-agent model is broad, it is still too narrow to serve as a general model of intelligent systems.

5.4 Intelligent systems need not be psychomorphic

What would count as high-level yet non-psychomorphic intelligence? One would be inclined to say that we have general, high-level AI if a coordinated pool of AI resources could, in aggregate:

- Do theoretical physics and biomedical research[1]
- Provide a general-purpose conversational interface[2] for discussing AI tasks
- Discuss and implement[3] designs for self-driving cars, spacecraft, and AI systems[4]
- Effectively automate development[5] of next-generation AI systems for AI design

None of these AI tasks is fundamentally different from translating languages, learning games, driving cars, or designing neural networks—tasks performed by systems not generally regarded as mind-like. Regarding the potential power such a coordinated pool of AI services, note that automating the development of AI systems for AI design[6] enables what amounts to recursive improvement.

5.5 Engineering and biological evolution differ profoundly

Rather than regarding artificial intelligence as something that fills a mind-shaped slot, we can instead consider AI systems as products of increasingly-automated technology development, an extension of the R&D process that we see in the world today. This development-oriented perspective on AI technologies highlights profound and pervasive differences between evolved and engineered intelligent systems (see table).

5.6 Studies of AI prospects have often made tacitly biological assumptions

Although technical models of artificial intelligence avoid overtly biological assumptions, it is nonetheless common (though far from universal!) to assume that advanced AI systems will:

1. Section 28: Automating biomedical R&D does not require defining human welfare
2. Section 21: Broad world knowledge can support safe task performance
3. Section 23: AI development systems can support effective human guidance
4. Section 12: AGI agents offer no compelling value
5. Section 10: R&D automation dissociates recursive improvement from AI agency
6. Section 10: R&D automation dissociates recursive improvement from AI agency

Organization of units:	Distinct organisms	Systems of components
Origin of new capabilities:	Incremental evolution	Research, development
Origin of instances:	Birth, development	Downloading files
Basis for learning tasks:	Individual experience	Aggregated training data
Transfer of knowledge:	Teaching, imitation	Copying models, parameters
Necessary competencies:	General life skills	Specific task performance
Success metric:	Reproductive fitness	Fitness for purpose
Self-modification:	Necessary	Optional
Continuity of existence:	Necessary	Optional
World-oriented agency:	Necessary	Optional

- Exist as individuals, rather than as systems of coordinated components[1]
- Learn from individual experience, rather than from aggregated training data[2]
- Develop through self-modification, rather than being constructed[3] and updated[4]
- Exist continuously, rather than being instantiated on demand[5]
- Pursue world-oriented goals, rather than performing specific tasks[6]

These assumptions have recognizable biological affinities, and they invite further assumptions that are tacitly biomorphic, psychomorphic, and even anthropomorphic.

5.7 Potential mind-like systems are situated in a more general space of potential intelligent systems

It has been persuasively argued that rational, mind-like superintelligence is an attractor in the space of potential AI systems, whether by design or inadvertent emergence. A crucial question, however, is the *extent of the basin of attraction* for mind-like systems within the far more general space of potential AI systems. The discussion above suggests that this basin is far from coextensive with the space of highly-capable AI systems, including systems that can, in aggregate, provide superintelligent-level services across an indefinitely

1. Section 15: Development-oriented models align with deeply-structured AI systems
2. Section 16: Aggregated experience and centralized learning support AI-agent applications
3. Section 21: Broad world knowledge can support safe task performance
4. Section 33: Competitive AI capabilities will not be boxed
5. Section 33: Competitive AI capabilities will not be boxed
6. Section 10: R&D automation dissociates recursive improvement from AI agency

wide range of tasks.[1] We cannot chart the space of potential AI problems and solutions solely within the confines of rational-agent models, because most of that space lies outside.

Further Reading

- *Section 2: Standard definitions of "superintelligence" conflate learning with competence*
- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 12: AGI agents offer no compelling value*
- *Section 15: Development-oriented models align with deeply-structured AI systems*

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

> 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

Although a common story suggests that any system composed of rational, high-level AI agents should (or must?) be regarded as a single, potentially powerful agent, the case for this idea is extraordinarily weak. AI service providers can readily satisfy the conditions for VNM rationality while employing knowledge and reasoning capacity of any level or scope. Bostrom's Orthogonality Thesis implies that even VNM-rational, SI-level agents need not maximize broad utility functions, and as is well known, *systems composed of rational agents* need not maximize any utility function at all. In particular, systems composed of competing AI service providers cannot usefully be regarded as unitary agents, much less as a unitary, forward-planning, utility-maximizing AGI agent. If, as seems likely, much of the potential solution-space for AI safety requires affordances like those in the AI-services model, then we must reconsider

1. Section 12: AGI agents offer no compelling value

long-standing assumptions regarding the dominance of utility-maximizing AI agents.

6.2 Systems of SI-level agents have been assumed to act as a single agent

In informal discussions of AI safety, it been widely assumed that, when considering a system comprising rational, utility-maximizing AI agents, one can (or should, or even *must*) model them as a single, emergent agent. This assumption is mistaken, and worse, impedes discussion of a range of potentially crucial AI safety strategies. To understand how we could employ and manage systems of rational agents, we can (without loss of generality) start by considering individual systems ("service providers") that act as rational agents.

6.3 Individual service providers can be modeled as individual agents

The von Neumann-Morgenstern expected-utility theorem shows that, if an agent meets a set of reasonable conditions defining rational behavior, the agent must choose actions that maximize the expected value of some function that assigns numerical values (utilities) to potential outcomes. If we consider a system that provides a service to be "an agent", then it is at least reasonable to regard VNM rationality as a condition for optimality.

6.4 Trivial agents can readily satisfy the conditions for VNM rationality

To *manifestly violate* the conditions for VNM rationality, an agent must make choices that are incompatible with any possible utility function. Accordingly, VNM rationality can be a trivial constraint: It is compatible, for example, with a null agent (that takes no actions), with an indifferent agent (that values all outcomes equally), and with any agent that acts only once (and hence cannot exhibit inconsistent preferences). Even among non-trivial agents, VNM rationality need not have deep or complex implications for world-oriented behaviors. Unlike humans,[1] computational systems do not necessarily (or

1. Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame.

even by default) act in an open world,[1] or value the "receipt" of "rewards",[2] or accumulate state changes over time,[3] or pursue outcomes beyond the completion of an immediate, atomic task.[4]

6.5 Trivially-rational agents can employ reasoning capacity of any scope

Predictive models are services that can provide building blocks for active AI services. Potential predictions include:

Text, language	→	Predicted human translation[5]
Present state	→	Predicted state
Agent, State	→	Predicted action
Action, State	→	Predicted outcome
Outcome	→	Predicted human approval[6]

The archetypical predictive model acts as a fixed function (*e.g.*, a translator of type T :: string → string). In each instance above, greater knowledge and reasoning capacity can improve performance; and in each instance, "actions" (function applications) may be judged by their external instrumental value, but cannot themselves violate the conditions for VNM rationality. As suggested by the examples above, fixed predictive models can encapsulate intelligence for active applications: For example, a system might drive a car to a user-selected destination while employing SI-level resources that inform steering decisions by predicting human (dis)approval of predicted outcomes.

6.6 High intelligence does not imply optimization of broad utility functions

Bostrom's (2014, p.107) Orthogonality Thesis states that "more or less any level of intelligence can be combined with more or less any final goal", and it follows that high-level intelligence can be applied tasks of bounded scope and duration that do not engender convergent instrumental goals.[7] Note that

1. Section 20: Collusion among superintelligent oracles can readily be avoided
2. Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents
3. Section 2: Standard definitions of "superintelligence" conflate learning with competence
4. Section 19: The orthogonality thesis undercuts the generality of instrumental convergence
7. Section 19: The orthogonality thesis undercuts the generality of instrumental convergence

service-*providers* need not seek to expand their capabilities: This is a task for service *developers*, while the service-developers are themselves *providers* of service-development services (the implied regress is loopy, not infinite).

6.7 Systems composed of rational agents need not maximize a utility function

There is no canonical way to aggregate utilities over agents, and game theory shows that interacting sets of rational agents need not achieve even Pareto optimality. Agents can compete to perform a task, or can perform adversarial tasks such as proposing and criticizing actions;[1] from an external client's perspective, these uncooperative interactions are features, not bugs (consider the growing utility of generative adversarial networks[2]). Further, *adaptive* collusion can be cleanly avoided: Fixed functions, for example, cannot negotiate or adapt their behavior to align with another agent's purpose.

In light of these considerations, it would seem strange to think that sets of AI services (even SI-level services) would necessarily or naturally collapse into utility-maximizing AI agents.

6.8 Multi-agent systems are *structurally* inequivalent to single agents

There is, of course, an even more fundamental objection to drawing a boundary around a set of agents and treating them as a single entity: In interacting with a set of agents, one can choose to communicate with one or another (*e.g.*, with an agent or its competitor); if we assume that the agents are in effect a single entity, we are assuming a constraint on communication that does not exist in the multi-agent model. The models are fundamentally, structurally inequivalent.

6.9 Problematic AI services need not be problematic AGI agents

Because AI services can in principle be fully general, combinations of services could of course be used to implement complex agent behaviors up to and including those of unitary AGI systems. Further, because evolutionary pressures can engender the emergence of powerful agents from lesser cognitive systems

1. Section 20: Collusion among superintelligent oracles can readily be avoided
2. `https://scholar.google.co.uk/scholar?as_ylo=2017&q=generative+adversarial+networks&hl=en&as_sdt=0,5`

(*e.g.*, our evolutionary ancestors), the unintended emergence of problematic agent behaviors in AI systems must be a real concern.[1]

Unintended, perverse interactions among some sets of service-providers will likely be a routine occurrence, familiar to contemporaneous researchers as a result of ongoing experience and AI safety studies.[2] Along this development path, the implied threat model is quite unlike that of a naïve humanity abruptly confronting a powerful, world-transforming AGI agent.

6.10 The AI-services model expands the solution-space for addressing AI risks

The AI-services and AGI-agent models of superintelligence are far from equivalent, and the AI-services model offers a wider range of affordances for structuring AI systems. Distinct, stable predictive models of human approval,[3] together with natural applications of competing and adversarial AI services,[4] can provide powerful tools for addressing traditional AI safety concerns. If, as seems likely, much of the potential solution-space for addressing AI x-risk[5] requires affordances within the AI-services model, then we must reconsider long-standing assumptions regarding the dominant role of utility-maximizing agents, and expand the AI-safety research portfolio to embrace new lines of inquiry.

Further Reading

- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*
- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*
- *Section 14: The AI-services model brings ample risks*
- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 19: The orthogonality thesis undercuts the generality of instrumental convergence*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 32: Unaligned superintelligent agents need not threaten world stability*

---

1. Section 14: The AI-services model brings ample risks
2. Section 35: Predictable aspects of future knowledge can inform AI safety strategies
3. Section 22: Machine learning can develop predictive models of human approval
4. Section 20: Collusion among superintelligent oracles can readily be avoided
5. Section 32: Unaligned superintelligent agents need not threaten world stability

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

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

Training agents on ill-defined human tasks may seem to be in conflict with developing distinct services provided by agents with bounded goals. Perceptions of conflict, however, seem rooted in anthropomorphic intuitions regarding connections between human-like skills and human-like goal structures, and more fundamentally, between learning and competence. These considerations are important to untangle because human-like training is arguably necessary to the achievement of important goals in AI research and applications, including adaptive physical competencies and perhaps general intelligence itself. Although performing safely-bounded tasks by applying skills learned through loosely-supervised exploration appears tractable, human-like world-oriented learning nonetheless brings unique risks.

7.2 Does training on human-like tasks conflict with the AI-services model?

Discussions suggest that many researchers see training on human-like tasks as a critical goal that is potentially in conflict with the AI-services model of general intelligence. Human-like tasks in natural environments (whether real or simulated) are often ill-defined and open-ended; they call for exploration of possibilities and creative planning that goes beyond traditional reinforcement learning based on clearly-defined reward functions. The AI-services model, by contrast, emphasizes the role of focused skills applied to bounded goals aligned with human purposes.

An examination of the distinction between human-like skills and human-like goal structures reduces the this apparent conflict. Critical differences emerge through the distinction between *intelligence as learning* and *intelligence as competence*,[1], the power of development architectures based on aggregated

1. Section 2: Standard definitions of "superintelligence" conflate learning with competence

experience and centralized learning,[1] and the role of application architectures in which tasks are naturally bounded.[2]

7.3 Human-like learning may be essential to developing general intelligence

General intelligence, in the sense of *general competence*, obviously includes human-like world-oriented knowledge and skills, and it would be surprising if these could be gained without human-like world-oriented learning. The practical value of human-like abilities is a major driving force behind AI research.

A more interesting question is whether human-like learning from human-like experience is essential to the development of general intelligence in the sense of *general learning ability*. This is a plausible hypothesis: Human intelligence is the only known example of what we consider to be general intelligence, and it emerged from open-ended interaction of humans with the world over the time spans of genetic, cultural, and individual development. Some researchers aim to reproduce this success by imitation.

Beyond the practical value of human-like learning, and the hypothesis that is may be essential to the development of general intelligence, there is a third reason to expect AI research to continue in this direction: Since Turing (1950), AI researchers have defined their objectives in terms of matching human capabilities in a broad sense, and have looked toward human-like learning, starting with child-like experience, as a natural path toward this goal.

7.4 Current methods build curious, imaginative agents

In pursuit of AI systems that can guide agents in complex worlds (to date, usually simulated), researchers have developed algorithms that build abstracted models of the world and use these as a basis for "imagination" to enable planning, reducing the costs of learning from trial and error (Weber et al. 2017; Nair et al. 2018; Ha and Schmidhuber 2018; Wayne et al. 2018). Reinforcement-learning agents have difficulty learning complex sequences of actions guided only by end-state goals; an effective approach has been development of algorithms that have "curiosity", seeking novel experiences

1. Section 16: Aggregated experience and centralized learning support AI-agent applications
2. Section 38: Broadly-capable systems coordinate narrower systems

and exploring actions of kinds that might be relevant to any of a range of potential goals (Pathak et al. 2017; Burda et al. 2018). Search guided by curiosity and imagination can adapt to novel situations and find unexpected solutions. Agents can also learn by observation of the real world, whether guided by demonstrations offered by human beings, or by imitating actions in relevant videos downloaded from the internet (Duan et al. 2017; Peng et al. 2018).

7.5 Human-like competencies do not imply human-like goal structures

Human beings learn human goal structures, but full human goal structures—life goals, for example—do not emerge directly or naturally from applying curiosity, imagination, and imitation to learning even an unbounded range of bounded tasks. The idea that human-like problem-solving is tightly linked to human-like goals may stem from what are tacitly biological background assumptions,[1] including evolution under competition, learning from individual experience, unitary functionality, and even physical continuity. The abstraction of AI systems as rational utility-directed agents also proves, on examination, to draw on tacitly anthropomorphic assumptions.[2]

Even speaking of "agents learning", as does the preceding section, is subject to this criticism. Humans learn by acting, but standard "reinforcement learning" methods sever this link: "Reinforcement learning" means training by a reinforcement learning algorithm, yet this algorithm performs no actions, while a trained agent learns nothing by acting.[3] Learning and action can of course be fused in a single system, yet they need not be, and learning can be more effective when they are separate.[4] Again, it is a mistake to conflate *intelligence as learning capacity* with *intelligence as competence*.[5]

---

1. Section **??**: **??**
2. Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame
3. Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents
4. Section 16: Aggregated experience and centralized learning support AI-agent applications
5. Section 2: Standard definitions of "superintelligence" conflate learning with competence

7.6 Open-ended learning can develop skills applicable to bounded tasks

In light of these considerations, it is natural and not necessarily problematic to pursue systems with general, human-like learning abilities as well as systems with collectively-general human-like competencies. Open-ended learning processes and general competencies do not imply problematic goals—and in particular, do not necessarily engender convergent instrumental goals.[1]

Even if the development of general intelligence though open-ended learning-to-learn entailed the development of opaque, unitary, problematic agents, their capabilities could be applied to the development of compact systems[2] that retain general learning capabilities while lacking the kinds of information, competencies, and goals that fit the profile of a dangerous AGI agent. Note that extracting and refining skills from a trained system can be a less challenging development task than learning equivalent skills from experience alone.

7.7 Human-like world-oriented learning nonetheless brings unique risks

These considerations suggest that there need be no direct line from human-like competencies to problematic goals, yet some human-like competencies are more problematic than, for example, expertise tightly focused on theorem proving or engineering design. Flexible, adaptive action in the physical world can enable disruptive competition in arenas that range from industrial parks to the battlefields. Flexible, adaptive action in the world of human information can support interventions that range from political manipulation to sheer absorption of human attention; the training objectives of the chatbot XiaoIce, for example, include engaging humans in supportive emotional relationships while maximizing the length of conversational exchanges (Zhou et al. 2018). She does this very well, and is continuing to learn.

Adherence to the AI services model by no means guarantees benign behaviors or favorable world outcomes,[3] even when applied with good intentions. Because of their potential for direct engagement with the world, however, human-like learning and capabilities present a special range of risks.

---

1. Section 19: The orthogonality thesis undercuts the generality of instrumental convergence
2. Section 30: Risky AI can help develop safe AI
3. Section 14: The AI-services model brings ample risks

Further Reading

- *Section 2: Standard definitions of "superintelligence" conflate learning with competence*
- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 14: The AI-services model brings ample risks*
- *Section 16: Aggregated experience and centralized learning support AI-agent applications*
- *Section 17: End-to-end reinforcement learning is compatible with the AI-services model*
- *Section 21: Broad world knowledge can support safe task performance*

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

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

Strong "optimization power" has often been assumed to increase AI risks by increasing the scope of a system's capabilities, yet task-focused optimization can have the opposite effect. Optimization of any system for a task constrains its structure and behavior, implicitly constraining its off-task capabilities: A competitive race car cannot transport a load of passengers, and a bus will never set a land speed record. In an AI context, optimization will tend to constrain capabilities and decrease risks when objectives are bounded in space, time, and scope, and when objective functions assign costs to both resource consumption and off-task effects. Fortunately, these are natural conditions for AI services. Optimizing AI systems for bounded tasks is itself a bounded task, and some bounded tasks (*e.g.*, predicting human approval) can contribute to general AI safety. These considerations indicate that strong, even SI-level optimization can both improve and constrain AI performance.

8.2 Strong optimization power need not increase AI capability and risk

"Optimization power" is widely regarded as a source of both AI capabilities and risks, but this concern is usually expressed in the context of open-ended objectives pursued with weak resource constraints. For tasks with bounded, cost-sensitive objectives, however, increasing optimization can have the opposite effect.

8.3 Strong optimization is a strong constraint

Full optimization of a system with respect to a value function typically yields not only a unique, maximal expected value, but robust constraints on the system itself. In practice, even approximate optimization strongly constrains both the structure and behavior of a system, thereby constraining its capabilities. In physical engineering, a race car cannot transport a load of passengers, and a bus will never set a land speed record; in AI development, an efficient text-translation system will never plan a vehicle path, and an efficient path-planning system will never provide translation services. Because off-task capabilities are costly, they will be excluded by cost-sensitive optimization.

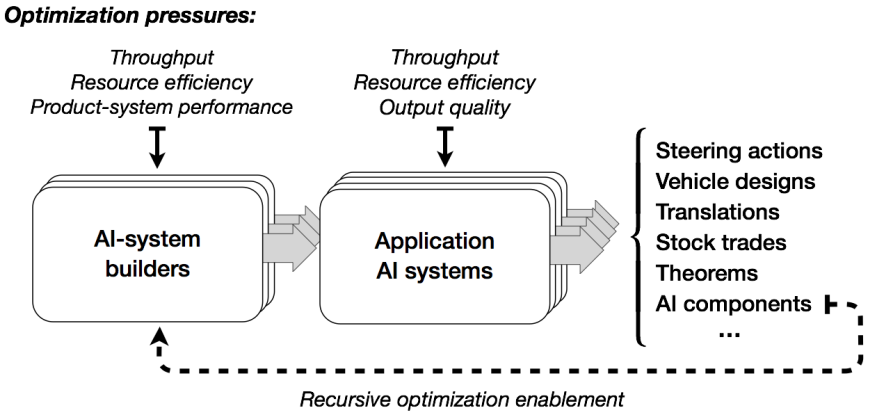

Figure 2: *Optimization pressures during AI system development focus resources on tasks and enable further development based on task-focused components.*

8.4 Optimization of AI systems can reduce unintended consequences

Optimization will tend to reduce risks when task objectives are bounded in space, time, and scope, and when the value-function assigns costs to both resource use and unintended human-relevant effects. With bounded objectives, remote and long-term effects will contribute no value and hence will be unintended, not actively optimized, and likewise for off-task consequences that are local and near-term. Going further, when costs are assessed for resource consumption and unintended, human-relevant effects,[1] stronger optimization will tend to actively reduce unwanted consequences (see Armstrong and Levinstein [2017] for a concept of *reduced-impact AI*).

8.5 Strong external optimization can strongly constrain internal capabilities

If the costs of appropriate computational resources are given substantial weight, strong optimization of an AI system can act as a strong constraint on its inputs and model capacity, and on the scope of its mechanisms for off-task inference, modeling, and planning. The weighting of computational-cost components need not reflect external economic costs, but can instead be chosen to shape the system under development.

A major concern regarding strong AI capabilities is the potential for poorly-defined goals and powerful optimization to lead to perverse plans that (for example) act through unexpected mechanisms with surprising side-effects. Optimization to minimize a system's off-task information, inference, modeling, and planning, however, can constrain the scope for formulating perverse plans, because the planning itself may incur substantial costs or require resources that have been omitted in the interest of efficiency. Optimization for on-task capabilities can thereby avoid a range of risks that have never been considered.

8.6 Optimizing an AI system for a bounded task is itself a bounded task

Optimizing a system for a bounded AI task can itself be framed as a bounded AI task. The task of designing and optimizing an AI system for a given task using given machine resources within an acceptably short time does not entail open-ended, world-affecting activity, and likewise for designing and

1. Section 22: Machine learning can develop predictive models of human approval

optimizing AI systems for the tasks involved in designing and optimizing AI systems for the tasks of designing and optimizing AI systems.

8.7 Superintelligent-level optimization can contribute to AI safety

Given that optimization can be used to shape AI systems that perform a range of tasks with little or no catastrophic risk, it may be useful to seek tasks that, in composition with systems that perform other tasks, directly reduce the risks of employing systems with powerful capabilities. A leading example is the development of predictive models of human relevance and human approval based on large corpora of human opinions and crowd-sourced challenges. Ideally, such models would have access to general world knowledge and be able to engage in general reasoning about cause, effect, and the range of potential human reactions. Predictive models of human approval[1] would be useful for augmenting human oversight,[2] flagging potential concerns,[3] and constraining the actions of systems with different capabilities.[4] These and similar applications are attractive targets for shaping AI outcome through differential technology development.

Bad actors could of course apply strongly optimized AI technologies—even approval modeling—to bad or risky ends (*e.g.*, open-ended exploitation of internet access for wealth maximization). Bad actors and bad actions are a crucial concern in considering strategies for managing a safe transition to a world with superintelligent-level AI, yet effective countermeasures may themselves require strong, safe optimization of AI systems for strategically important tasks. The development of strong optimization power is a given, and we should not shy away from considering how strongly optimized AI systems might be used to solve problems.

Further Reading

- *Section 2: Standard definitions of "superintelligence" conflate learning with competence*
- *Section 12: AGI agents offer no compelling value*

---

1. Section 22: Machine learning can develop predictive models of human approval
2. Section 24: Human oversight need not impede fast, recursive AI technology improvement
3. Section 20: Collusion among superintelligent oracles can readily be avoided
4. Section 20: Collusion among superintelligent oracles can readily be avoided

- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 30: Risky AI can help develop safe AI*

9 Opaque algorithms are compatible with functional transparency and control

> 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

Deep-learning methods employ opaque algorithms that operate on opaque representations, and it would be unwise to assume pervasive transparency in future AI systems of any kind. Fortunately, opacity at the level of algorithms and representations is compatible with transparency at higher levels of system functionality. We can shape information inputs and training objectives at component boundaries, and can, if we choose, also shape and monitor information flows among opaque components in larger systems. Thus, substantial high-level understanding and control is compatible with relaxed understanding of internal algorithms and representations. As always, the actual *application* of potential control measures can be responsive to future experience and circumstances.

9.2 Deep-learning methods are opaque and may remain so

The products of deep learning are notoriously opaque: Numerical transformations produce numerical vectors that can be decoded into useful results, but the encodings themselves are often incomprehensible. Opacity is the norm, and interpretability is the exception. In considering problems of AI control and safety, it would be unwise to assume pervasive transparency.

9.3 The scope of information and competencies can be fuzzy, yet bounded

Although we may lack knowledge of how a deep learning system represents information and algorithms, we can have substantial knowledge of the scope of its information and competencies. For example, information that is absent

from a system's inputs (in both training and use) will be absent from its algorithms, state, and outputs. Inference capabilities may blur the scope of given information, but only within limits: A Wikipedia article cannot be inferred from language-free knowledge of physics. Likewise, while the scope of a system's competencies may be fuzzy, competencies far from a system's task focus (*e.g.*, theorem-proving competencies in a vision system, or vehicle-guidance competencies in a language-translation system) will be reliably absent. Bounds on information and competencies are natural and inevitable, and can be applied to help us understand and constrain AI-system functionality.

9.4 Restricting resources and information at boundaries constrains capabilities

Several obvious affordances for control are available at the boundaries of AI systems. For example, tasks that require absent information cannot be performed, and the distinct role of physical memory in digital systems enables a clean separation of episodic task performance from cumulative learning.[1] Controls at boundaries have transitive effects within systems of collaborating components: A component cannot transfer information that it does not have, regardless of how internal communications are encoded.

Further, competitive systems must deliver results in bounded times and with bounded resources. Optimization pressures (*e.g.*, on model capacity, training time, and execution cost) will tend to exclude investments in off-task capacities and activities, and stronger optimization will tend to strengthen, not weaken, those constraints.[2] A system trained to provide services to other systems might perform unknown tasks, yet those tasks will not be *both* costly and irrelevant to external objectives.

These considerations are fundamental: They apply regardless of whether an AI system is implemented on digital, analog, or quantum computational hardware, and regardless of whether its algorithms are neural and trained, or symbolic and programmed. They scale to task domains of any scope, and to systems of any level of intelligence and competence.

---

1. Section 2: Standard definitions of "superintelligence" conflate learning with competence
2. Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects

9.5 Providing *external* capabilities can constrain *internal* capabilities

Under appropriate optimization pressures, a system trained with access to an efficient resource with particular capabilities[1] will not itself develop equivalent capabilities, and use of those particular capabilities will then involve use of an identifiable resource. This mechanism provides an affordance for shaping the organization of task-relevant capabilities in the development of piecewise-opaque systems. Potential advantages include not only functional transparency, but opportunities to ensure that components (vision systems, physical models, *etc.*) are well-trained, well-tested, and capable of good generalization within their domains.

9.6 Deep learning can help interpret internal representations

Deep learning techniques can sometimes provide insight into the content of opaque, learned representations. To monitor the presence or absence of a particular kind of information in an opaque (but not *adversarially* opaque) representation, deep learning can be applied to attempt to extract and apply that information. For example, a representation may be opaque to humans, but if it supports an image-recognition task, then the representation must contain image information; if not, then it likely doesn't.

9.7 Task-space models can enable a kind of "mind reading"

The task-space model[2] of general intelligence suggests that the subtasks engaged by problem-solving activities can (both in principle and in practice) be associated with regions in semantic spaces. Different high-level tasks will generate different footprints of activity in the space of subtasks, and one need not understand how every subtask is represented or performed to understand *what the task is about.*

Restricting the range of task-space accessible to a system could potentially provide a mechanism for constraining its actions, while observing access patterns could potentially provide the ability to monitor the considerations that go into a particular action. For example, it would be unremarkable for a system that organizes food production to access services applicable to food preparation and delivery, while a system that accesses services applicable to

1. Section 39: Tiling task-space with AI services can provide general AI capabilities
2. Section 39: Tiling task-space with AI services can provide general AI capabilities

synthesizing neurotoxins for delivery in food might trigger a warning. Such insights (a kind of "mind reading") could be useful both in advancing AI safety and in solving more prosaic problems of system development and debugging.

9.8 The application of control measures can be adapted to experience and circumstances

Whether any particular set of control measures should be applied, and to what extent, is a question best answered by the AI community as circumstances arise. Experience will provide considerable knowledge[1] of which kinds of systems are reliable, which fail, and which produce surprising (perhaps *disturbingly* surprising) results. Along the way, conventional concerns regarding safety and reliability will drive efforts to make systems better understood and more predictable.

To catalog a range of potential control measures (*e.g.*, attention to information content, task focus, and optimization pressures) is not to assert that any particular measure or intensity of application will be necessary or sufficient. The value of inquiry in this area is to explore mechanisms that *could* be applied in response to future experience and circumstances, and that may deserve attention today as safety-relevant components of general AI research.

Further Reading

- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*
- *Section 38: Broadly-capable systems coordinate narrower systems*
- *Section 39: Tiling task-space with AI services can provide general AI capabilities*

1. Section 35: Predictable aspects of future knowledge can inform AI safety strategies

10 R&D automation dissociates recursive improvement from AI agency

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

10.1 Summary

AI development processes can be automated by applying specialized AI competencies to AI-development tasks, and incremental automation of AI development systems can continue to reflect the familiar organization of research and development processes. Asymptotically recursive technology improvement requires neither self-improving components nor agents that act in the world, and can provide general AI functionality without recourse to general AI agents. The R&D-automation model offers a range of potential affordances for addressing AI safety concerns.

10.2 R&D automation can employ on diverse, specialized AI tools

AI research and development, like other areas of software technology, exploits a growing range of automated tools. As AI technologies approach or exceed human competence in AI development tasks, we can expect to see incremental automation throughout the development process, asymptotically enabling recursive technology improvement. The development-automation model differs from classic models of recursive improvement in that it does not call for self-improving general-purpose agents.

10.3 AI R&D automation will reflect universal aspects of R&D processes

The R&D process links the discovery of general principles and mechanisms to the construction of complex systems tailored to specific functions and circumstances. The R&D-automation model describes AI development today, and increasing automation of AI development seems unlikely to obliterate this deep and universal task structure.

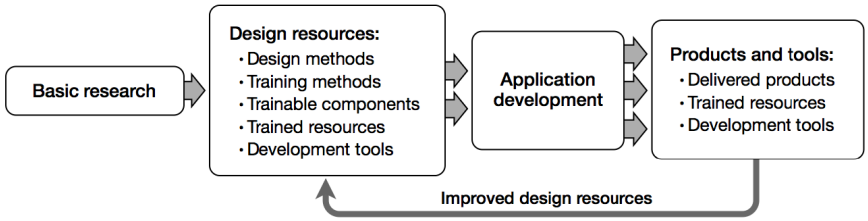

Figure 3: *Schematic model of the current AI research and development pipeline.*

10.4 AI R&D automation will reflect the structure of AI development tasks

In AI R&D, the elements of development tasks are organized along lines suggested by the following diagram:

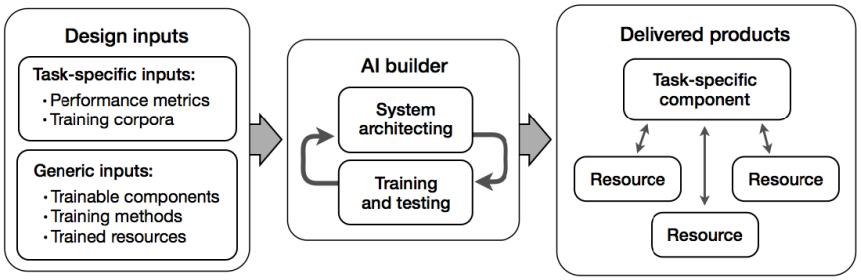

Figure 4: *Schematic model of an AI-enabled application-oriented system development task that draws on a range of previously developed components.*

In architecting AI systems, application developers can draw on relatively generic sets of trainable components and training methods, and can compose these with previously developed resources. The resulting architectures are trained with task-specific performance metrics and corpora, and revised based on results.

Incremental automation does not change the fundamental structure of development tasks: The R&D automation model accommodates AI-enabled innovation in components, methods, resources, training corpora, and performance metrics; the model also accommodates ongoing involvement of human developers in any role. Over time, we can expect human roles to shift from technical implementation to general task description and performance evaluation.

10.5 AI R&D automation leads toward recursive technology improvement

In the schematic diagram above, an AI builder is itself an R&D product, as are AI systems that perform exploratory AI research. Pervasive and asymptotically complete AI-enabled automation of AI R&D can enable what amounts to recursive improvement, raising the yield of AI progress per unit of human effort without obvious bound. In this model there is no locus of activity that corresponds to recursive "self" improvement; as we see in today's AI R&D community, loosely coupled activities are sufficient to advance all aspects of AI technology.

10.6 General SI-level functionality does not require general SI-level agents

The classic motivation for building self-improving general-purpose superintelligent agents is to provide systems that can perform a full range of tasks with superintelligent competence. The R&D-automation model, however, shows how to provide, on demand, systems that can perform any of a fully general range of tasks without invoking the services of a fully general agent.

10.7 The R&D-automation model reframes the role of AI safety studies and offers potential affordances for addressing AI safety problems

In the present framework, agent-oriented AI safety research plays the dual roles of expanding the scope of safe agent functionality and identifying classes of systems and applications (including tightly coupled configurations of R&D components) in which radically unsafe agent behaviors might arise unintentionally. In other words, agent-oriented safety work can both find safe paths and mark potential hazards.

The R&D-automation model describes component-based systems that are well-suited to the production of component-based systems, hence it invites consideration of potential safety-relevant affordances of deeply-structured AI implementations. In particular, the prospect of safe access to superintelligent machine learning invites consideration of predictive models of human approval that are trained on large corpora of human responses to events and actions, and subsequently serve as components of structured, approval-directed AI agents.

Further Reading

- *Section 12: AGI agents offer no compelling value*
- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 30: Risky AI can help develop safe AI*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*

11 Potential AGI-enabling technologies also enable comprehensive AI services

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

In catastrophic runaway-AI scenarios, systems capable of self-improvement lead to—and hence precede—opaque AGI agents with general superhuman competencies. Systems capable of self-improvement would, however, embody high-level development capabilities that could first be exploited to upgrade ongoing, relatively transparent AI R&D automation. Along this path, transparency and control need not impede AI development, and optimization pressures can sharpen task focus rather than loosen constraints. Thus, in scenarios where advances in technology would enable the implementation of powerful but risky AGI agents, those same advances could instead be applied to provide comprehensive AI services—and stable, task-focused agents—while avoiding the potential risks of self-modifying AGI-agents.

11.2 In runaway-AGI scenarios, self-improvement precedes risky competencies

Self-improving, general-purpose AI systems would, by definition, have the ability to build AI systems with capabilities applicable to AI development tasks. In classic AGI-takeover scenarios, the specific competencies that enable algorithmic self-improvement would precede more general superhuman

competencies in modeling the world, defining world-changing goals, and pursuing (for example) a workable plan to seize control. Strong AI development capabilities would precede potential catastrophic threats.

11.3 "Self"-improvement mechanisms would first accelerate R&D

In any realistic development scenario, highly-capable systems will follow less-capable predecessors (*e.g.*, systems with weaker architectures, smaller datasets, or less training), and developers will have practical knowledge of how to instantiate, train, and apply these systems. Along paths that lead to systems able to implement more capable systems with little human effort,[1] it will be natural for developers to apply those systems to specific development tasks. Developing and packaging an opaque, self-improving AI system might or might not be among those tasks.

11.4 "Self"-improvement mechanisms have no special connection to agents

The option to develop and apply self-improving AGI systems would be compelling only if there were no comparable or superior alternatives. Given AI systems able to implement a wide range of AI systems, however, there would be no compelling reason to package and seal AI development processes in an opaque box. Quite the opposite: Practical considerations generally favor development, testing, and integration of differentiated components.[2] Potential AGI-level technologies could presumably automate such processes, while an open system-development architecture would retain system-level transparency and process control.

11.5 Transparency and control need not impede the pace of AI development

Because the outputs of basic research—the building blocks of technology improvement—need not directly affect the world, the need for human intervention in basic research is minimal and need not impede progress.[3] Meanwhile, in applications, guiding development to serve human purposes is not a

1. Section 10: R&D automation dissociates recursive improvement from AI agency
2. Section 15: Development-oriented models align with deeply-structured AI systems
3. Section 24: Human oversight need not impede fast, recursive AI technology improvement

burden, but an inherent part of the task of providing beneficial AI services.[1] Note that developing and applying AI systems to help humans guide development is itself a task within the scope of comprehensive AI R&D automation.[2]

11.6 Optimization pressures sharpen task focus

Thinking about prospects of applying high-level AI capabilities to the design and optimization of AI systems has been muddied by the tacit assumption that optimizing performance implies relaxing constraints on behavior. For any bounded task, however, this is exactly wrong: The stronger the optimization, the stronger the constraints.[3] In the context of AI-enabled AI-development, effective optimization of a development system will tend to minimize resources spent on off-task modeling and search. Regardless of their internal complexity, optimized components of AI development systems will have no spare time to daydream about world domination.

11.7 Problematic emergent behaviors differ from classic AGI risks

Systems of optimized, stable components can be used to implement fully general mechanisms, hence some configurations of components could exhibit problematic emergent behaviors of unexpected kinds. We can expect and encourage developers to note and avoid architectures of the kind that produce unexpected behaviors,[4] perhaps aided by AI-enabled analysis of both AI objectives,[5] and proposed implementations.[6] Avoiding problematic emergent behaviors in task-oriented systems composed of stable components is inherently more tractable than attempting to confine or control a self-modifying AGI system that might by default act as superintelligent adversarial agent.

---

1. Section 23: AI development systems can support effective human guidance
2. Section 23: AI development systems can support effective human guidance and Section 22: Machine learning can develop predictive models of human approval
3. Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects
4. Section 35: Predictable aspects of future knowledge can inform AI safety strategies
5. Section 22: Machine learning can develop predictive models of human approval
6. Section 26: Superintelligent-level systems can safely provide design and planning services

11.8 Potential AGI technologies might best be applied to automate development of comprehensive AI services

In summary, an AI technology base that could implement powerful self-improving AGI agents could instead be applied to implement (or more realistically, upgrade) increasingly automated AI development, a capability that in turn can be applied to implement a comprehensive range of AI applications. Thus, swift, AI-enabled improvement of AI technology does not require opaque self-improving systems,[1] and comprehensive AI services need not be provided by potentially risky AGI agents.[2]

Further Reading

- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 12: AGI agents offer no compelling value*
- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 30: Risky AI can help develop safe AI*
- *Section 33: Competitive AI capabilities will not be boxed*

12 AGI agents offer no compelling value

> 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

Practical incentives for developing AGI agents appear surprisingly weak. Providing comprehensive AI services calls for diverse, open-ended AI capabilities (including stable agent services), but their development does not require agents in any conventional sense. Although both the AGI and AI-service models can deliver general capabilities, their differences have a range of consequences; for example, by enabling access to stable AI components, competing

1. Section 10: R&D automation dissociates recursive improvement from AI agency
2. Section 12: AGI agents offer no compelling value

implementations, and adversarial checking mechanisms, the CAIS model offers safety-relevant affordances that the classic AGI model does not. Both the CAIS and AGI models propose recursive improvement of AI technologies, yet they differ in their accessibility: While CAIS models anticipate accelerating R&D automation that extends conventional development methodologies, AGI models look toward conceptual breakthroughs to enable self-improvement and subsequent safe application. Because AI development services could be used to implement AGI agents, the CAIS model highlights the importance of classic AGI-safety problems, while access to SI-level services could potentially mitigate those same problems.

12.2 Would AGI development deliver compelling value?

It is widely believed that the quest to maximize useful AI capabilities will necessarily culminate in artificial general intelligence (AGI), which is taken to imply AI agents that would be able to self-improve to a superintelligent level, potentially gaining great knowledge, capability, and power to influence the world. It has been suggested AGI may be effectively unavoidable either because:

1. Self-improving AI may almost unavoidably generate AGI agents, or
2. AGI agents would provide unique and compelling value, making their development almost irresistibly attractive.

However, the non-agent-based R&D automation dissociates recursive improvement from AI agency undercuts claim (1), while the prospective result of ongoing R&D automation, general AI development services, undercuts claim (2).

12.3 AI systems deliver value by delivering services

In practical terms, we value potential AI systems for what they could do, whether driving a car, designing a spacecraft, caring for a patient, disarming an opponent, proving a theorem, or writing a symphony. Scientific curiosity and long-standing aspirations will encourage the development of AGI agents with open-ended, self-directed, human-like capabilities, but the more powerful drives of military competition, economic competition, and improving human welfare do not in themselves call for such agents. What matters in practical terms are the concrete AI services provided (their scope, quality, and reliability) and the ease or difficulty of acquiring them (in terms of time, cost, and human effort).

12.4 Providing diverse AI services calls for diverse AI capabilities

Diverse AI services resolve into diverse tasks, some shared across many domains (*e.g.*, applying knowledge of physical principles, of constraints on acceptable behavior, *etc.*), while other tasks are specific to a narrower range of domains. Reflection on the range of potential AI services (driving a car, proving a theorem...) suggests the diversity of underlying AI tasks and competencies. We can safely assume that:

- No particular AI service will require all potential AI competencies.
- Satisfying general demands for new AI services will require a general ability to expand the scope of available competencies.

12.5 Expanding AI-application services calls for AI-development services

Developing a new AI service requires understanding its purpose (guided by human requests, inferred preferences, feedback from application experience, *etc.*), in conjunction with a process of design, implementation, and adaptation that produces and improves the required AI capabilities. Capability-development tasks can be cast in engineering terms: They include function definition, design and implementation, testing and validation, operational deployment, in-use feedback, and ongoing upgrades.

12.6 The AGI and CAIS models organize similar functions in different ways

In the CAI-services model, capability-development functions are explicit, exposed, and embodied in AI system components having suitable capacities and functional relationships. In the CAIS model, AI-enabled products are distinct from AI-enabled development systems, and the CAIS model naturally emerges from incremental R&D automation.

In the classic AGI-agent model, by contrast, capability-development functions are implicit, hidden, and embodied in a single, conceptually-opaque, self-modifying agent that pursues (or is apt to pursue) world-oriented goals. Thus, capability development is internal to an agent that embodies both the development mechanism and its product. Implementation of the AGI model is widely regarded as requiring conceptual breakthroughs.

12.7 The CAIS model provides additional safety-relevant affordances

The CAIS model both exemplifies and naturally produces deeply structured AI systems[1] based on identifiable, functionally-differentiated components. Structured architectures provide affordances for both component- and system-level testing, and for the re-use of stable, well-tested components (*e.g.*, for vision, motion planning, language understanding. . .) in systems that are adapted to new purposes. These familiar features of practical product development and architecture can contribute to reliability in a conventional sense, but also to AI safety in the context of superintelligent-level competencies.

In particular, the CAIS model offers component and system-level affordances for structuring information inputs and retention, mutability and stability, computational resource allocation, functional organization, component redundancy, and internal process monitoring; these features distance the CAIS from opaque, self-modifying agents, as does the fundamental separation of AI products from AI development processes. In the CAIS context, components (*e.g.*, predictive models of human preferences) can be tested separately (or in diverse testbed contexts) without the ambiguities introduced by embedding similar functionality in systems with agent-level goals and potential incentives for deception.

Constraining AI systems through external, structural affordances:

Knowledge metering to bound information scope
Model distillation to bound information quantity
Checkpoint/restart to control information retention
Focused curricula to train task specialists
Specialist composition to address complex tasks
Optimization applied as a constraint

Figure 5: *AI development processes provide affordances for constraining AI systems that can be effective without insights into their internal representations. Points of control include information inputs, model size, (im)mutability, loss functions, functional specialization and composition, and optimization pressures that tend to become sharper as implementation technologies improve.* (Adapted from Drexler [2015])

1. Section 15: Development-oriented models align with deeply-structured AI systems

12.8 The CAIS model enables competition and adversarial checks

In developing complex systems, it is common practice to apply multiple analytical methods to a proposed implementation, to seek and compare multiple proposals, to submit proposals to independent review, and where appropriate, to undertake adversarial red-team/blue-team testing. Each of these measures can contribute to reliability and safety, and each implicitly depends on the availability of independent contributors, evaluators, testers, and competitors. Further, each of these essentially adversarial services scales to the superintelligent limit.

In the CAIS model, it is natural to produce diverse, independent, task-focused AI systems that provide adversarial services. By contrast, it has been argued that, in the classic AGI model, strong convergence (through shared knowledge, shared objectives, and strong utility optimization under shared decision theories) would render multiple agents effectively equivalent, undercutting methods that would rely on their independence. Diversity among AI systems is essential to providing independent checks, and can enable the prevention of potential collusive behaviors.[1]

12.9 The CAIS model offers generic advantages over classic AGI models

- *CAIS (like AGI)* encompasses recursive improvement of AI technologies, and hence could enable full-spectrum AI services that operate at a superintelligent level.
- *CAIS (but not AGI)* grows out of incremental R&D automation within the architecture of established development methodologies.
- *AGI (but not CAIS)* calls for conceptual breakthroughs to enable both implementation and subsequent safe application.
- *CAIS (but not AGI)* offers structural affordances for increasing reliability and safety, including diverse adversarial checks that scale to superintelligent systems.

12.10 CAIS affordances mitigate but do not solve AGI-control problems

Because systems that can implement AI functionality at a superintelligent level can presumably be used to implement classic AGI systems, CAI services would lower barriers to the development of AGI. Given the widespread

1. Section 20: Collusion among superintelligent oracles can readily be avoided

desire to realize the dream of AGI, it seems likely that AGI will, in fact, be realized unless actively prevented. Nonetheless, in a world potentially stabilized by security-oriented applications of superintelligent-level AI capabilities, prospects for the emergence of AGI systems may be less threatening. Superintelligent-level aid in understanding and implementing solutions to the AGI control problem[1] and could greatly improve our strategic position.

There is no bright line between safe CAI services and unsafe AGI agents, and AGI is perhaps best regarded as a potential branch from an R&D-automation/CAIS path. To continue along safe paths from today's early AI R&D automation to superintelligent-level CAIS calls for an improved understanding of the preconditions for AI risk, while for any given level of safety, a better understanding of risk will widen the scope of known-safe system architectures and capabilities.

The analysis presented above suggests that CAIS models of the *emergence of superintelligent-level AI capabilities*, including AGI, should be given substantial and arguably predominant weight in considering questions of AI safety and strategy.

Further Reading

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*

1. Section 20: Collusion among superintelligent oracles can readily be avoided

13 AGI-agent models entail greater complexity than CAIS

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

Recent discussions suggest that it would be useful to compare the relative complexities of AGI-agent and comprehensive AI services (CAIS) models of general intelligence. The functional requirements for open-ended self-improvement and general AI capabilities are the same in both instances, but made more difficult in classic AGI models, which require that fully-general functionality be internal to an autonomous, utility-directed agent. The rewards for accomplishing this compression of functionality are difficult to see. To attempt to encompass general human goals within the utility function of a single, powerful agent would reduce none of the challenges of aligning concrete AI behaviors with concrete human goals, yet would increase the scope for problematic outcomes. This extreme compression and its attendant problems are unnecessary: Task-oriented AI systems within the CAIS framework could apply high-level reasoning and broad understanding to a full spectrum of goals, coordinating open-ended, collectively-general AI capabilities to provide services that, though seamlessly integrated, need not individually or collectively behave as a unitary AGI agent.

13.2 Classic AGI models neither simplify nor explain self improvement

The classic AGI agent model posits open-ended self improvement, but this simple concept hides what by nature must be functionally equivalent to fully-automated and open-ended AI research and development. Hiding the complexity of AI development in a conceptual box provides only the illusion of simplicity. Discussion within the classic AGI model typically assumes an unexplained breakthrough in machine learning capabilities. For simplicity, an AI-services model could arbitrarily assume equivalent capabilities (perhaps based on the same hypothetical breakthrough), a deeper model offers a

framework for considering their implementation. To be comprehensive, AI services must of course include the service of developing new services, and current research practice shows that expanding the scope of AI services can be both incremental and increasingly automated.

13.3 Classic AGI models neither simplify nor explain general AI capabilities

Similarly, the classic AGI agent model posits systems that could provide general, fluidly-integrated AI capabilities, but this seemingly simple concept hides what by nature must be functionally equivalent to a comprehensive range of AI services and coordination mechanisms. The classic model assumes these capabilities without explaining how they might work; for simplicity, an AI-services model could arbitrarily assume equivalent capabilities, but a deeper model offers a framework for considering how diverse, increasingly comprehensive capabilities could be developed and integrated by increasingly automated means.

Note that, by intended definition, the "C" in CAIS is effectively equivalent to the "G" in AGI. Accordingly, to propose that an AGI agent could provide services beyond the scope of CAIS is either to *misunderstand* the CAIS model, or to *reject* it, *e.g.*, on grounds of feasibility or coherence. To be clear, fully realized CAIS services would include the service of coordinating and providing a seamless interface to other services, modeling behaviors one might have attributed to aligned AGI agents. The CAIS model of course extends to the provision of potentially dangerous services, including the service of building unaligned AGI agents.

13.4 Classic AGI models increase the challenges of AI goal alignment

The classic AGI model posits the construction of a powerful, utility-directed, superintelligent agent, a conceptual move that both engenders the problems of aligning a superintelligent agent's overall goals with human values and amalgamates and abstracts away the concrete problems that arise in aligning specific, useful behaviors with diverse and changing human goals. Although a simplified AI-services model could arbitrarily assume aligned systems with bounded goals and action spaces, a deeper model offers a framework for considering how such systems could be developed and how their development might go wrong—for example, by indirectly and inadvertently giving rise to agents with problematic goals and capabilities. Many of the questions first

framed as problems of AGI safety still arise, but in a different and perhaps more tractable systemic context.

13.5 The CAIS model addresses a range of problems without sacrificing efficiency or generality

To summarize, in each of the areas outlined above, the classic AGI model both obscures and increases complexity: In order for general learning and capabilities to fit a classic AGI model, they must not only exist, but must be integrated into a single, autonomous, self-modifying agent. Further, achieving this kind of integration would increase, not reduce, the challenges of aligning AI behaviors with human goals: These challenges become more difficult when the goals of a single agent must motivate all (and only) useful tasks.

Agent-services that are artificial, intelligent, and general are surely useful, both conceptually and in practice, but fall within the scope of comprehensive agent (and non-agent) AI services. The key contribution of the CAIS model is to show how integrated, fully-general AI capabilities could be provided within an open-ended architecture that is natural, efficient, relatively transparent, and quite unlike a willful, uniquely-powerful agent.

Further Reading

- *Section 1: R&D automation provides the most direct path to an intelligence explosion*
- *Section 6: A system of AI services is not equivalent to a utility maximizing agent*
- *Section 12: AGI agents offer no compelling value*
- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*

14 The AI-services model brings ample risks

> 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

Prospects for general, high-level AI services reframe—but do not eliminate—a range of AI risks. On the positive side, access to increasingly comprehensive AI services (CAIS) can reduce the practical incentives for developing

potentially problematic AGI agents while providing means for mitigating their potential dangers. On the negative side, AI services could facilitate the development of dangerous agents, empower bad actors, and accelerate the development of seductive AI applications with harmful effects. A further concern—avoiding perverse agent-like behaviors arising from interactions among service providers—calls for further study that draws on agent-centric models. Taking the long view, the CAIS model suggests a technology-agnostic, relatively path-independent perspective on potential means for managing SI-level AI risks.

14.2 Prospects for general, high-level AI services reframe AI risks

In a classic model of high-level AI risks, AI development leads to self-improving agents that gain general capabilities and enormous power relative to the rest of the world. The AI-services model[1] points to a different prospect: Continued automation of AI R&D[2] (viewed as an increasingly-comprehensive set of development services) leads to a general ability to implement systems that provide AI services, ultimately scaling to a superintelligent level. Prospects for comprehensive AI services (CAIS) contrast sharply with classic expectations that center on AGI agents: The leading risks and remedies differ in both nature and context.

14.3 CAIS capabilities could mitigate a range of AGI risks

On the positive side, capabilities within the CAIS model can be applied to mitigate AGI risks. The CAIS model arises naturally from current trends in AI development and outlines a more accessible path[3] to general AI capabilities; as a consequence, CAIS points to a future in which AGI agents have relatively low marginal instrumental value[4] and follow rather than lead the application of superintelligent-level AI functionality[5] to diverse problems. Accordingly, the CAIS model suggests that high-level agents will (or readily could) be developed in the context of safer, more tractable AI systems[6] that

1. Section 12: AGI agents offer no compelling value
2. Section 10: R&D automation dissociates recursive improvement from AI agency
3. Section 10: R&D automation dissociates recursive improvement from AI agency
4. Section 12: AGI agents offer no compelling value
5. Section 11: Potential AGI-enabling technologies also enable comprehensive AI services
6. Section 29: The AI-services model reframes the potential *roles* of AGI agents

can provide services useful for managing such agents. Predictive models of human concerns[1] are a prominent example of such services; others include AI-enabled capabilities for AI-systems design,[2] analysis,[3] monitoring, and upgrade.[4]

14.4 CAIS capabilities could facilitate the development of dangerous AGI agents

Comprehensive AI services necessarily include the service of developing useful AI agents with stable, bounded capabilities, but superintelligent-level CAIS could also be employed to implement general, autonomous, self-modifying systems that match the specifications for risky AGI agents. If not properly directed or constrained—a focus of current AI-safety research—such agents could pose catastrophic or even existential risks to humanity. The AI-services model suggests broadening studies of AI safety to explore potential applications of CAIS-enabled capabilities to risk-mitigating differential technology development, including AI-supported means for developing safe AGI agents.

14.5 CAIS capabilities could empower bad actors.

AI-service resources *per se* are neutral in their potential applications, and human beings can already apply AI services and products to do intentional harm. Access to advanced AI services could further empower bad actors in ways both expected and as-yet unimagined; in compensation, advanced AI services could enable detection and defense against bad actors. Prospective threats and mitigation strategies call for exploration and study of novel policy options.

14.6 CAIS capabilities could facilitate disruptive applications

Today's disruptive AI applications are services that serve some practical purpose. As shown by current developments, however, one person's service may be another person's threat, whether to business models, employment, privacy, or military security. Increasingly comprehensive and high-level AI services

1. Section 22: Machine learning can develop predictive models of human approval
2. Section 26: Superintelligent-level systems can safely provide design and planning services
3. Section 23: AI development systems can support effective human guidance
4. Section 16: Aggregated experience and centralized learning support AI-agent applications

will continue this trend, again in ways both expected and as-yet imagined. Mitigation of non-malicious disruption raises political, legal, and economic questions, because both the disruptors and the disrupted may have conflicting yet legitimate interests.

14.7 CAIS capabilities could facilitate seductive and addictive applications

Some disruptive applications provide seductive services that are detrimental, yet welcomed by their targets. AI systems today are employed to make games more addictive, to build comfortable filter bubbles, and to optimize message channels for appeal unconstrained by truth. There will be enormous scope for high-level AI systems to please the people they harm, yet mitigation of the individual and societal consequences of unconstrained seductive and addictive services raises potentially intractable questions at the interface of values and policy.

14.8 Conditions for avoiding emergent agent-like behaviors call for further study

Although it is important to distinguish between pools of AI services and classic conceptions of integrated, opaque, utility-maximizing agents, we should be alert to the potential for coupled AI services to develop emergent, unintended, and potentially risky agent-like behaviors. Because there is no bright line between agents and non-agents, or between rational utility maximization and reactive behaviors shaped by blind evolution, avoiding risky behaviors calls for at least two complementary perspectives: both (1) design-oriented studies that can guide implementation of systems that will provide requisite degrees of *e.g.*, stability, reliability, and transparency, and (2) agent-oriented studies support design by exploring the characteristics of systems that could display emergent, unintended, and potentially risky agent-like behaviors. The possibility (or likelihood) of humans implementing highly-adaptive agents that pursue open-ended goals in the world (*e.g.*, money-maximizers) presents particularly difficult problems.

Further Reading

- *Section 3: To understand AI prospects, focus on services, not implementations*
- *Section 7: Training agents in human-like environments can provide useful, bounded services*

- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 12: AGI agents offer no compelling value*

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

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

AI safety research has focused on unitary, relatively unstructured models of superintelligent systems. Development-oriented models, however, suggest that advanced AI systems will in practice comprise or employ structured systems of task-oriented components. Prospects for heterogeneous, deeply structured systems are best understood by considering development processes in which structure results from composing components, not from partitioning a hypothetical unitary functionality. A focus on development processes that lead to structured products links AI safety to ongoing AI research practice and suggests a range of topics for further inquiry.

15.2 AI safety research has often focused on unstructured rational-agent models

Research in AI safety has been motivated by prospects for recursive improvement that enables the rapid development of systems with general, superhuman problem-solving capabilities. Research working within this paradigm has centered not on the process of recursive improvement, but on its potential products, and these products have typically been modeled as discrete, relatively unstructured, general-purpose AI systems.

15.3 Structured systems are products of structured development

In an alternative, potentially complementary model of high-level AI, the products of recursive improvement are deeply structured, task-focused systems that collectively deliver a comprehensive range of superintelligent task capabilities, yet need not be (or become) discrete entities that individually span a

full range of capabilities. This model has been criticized as potentially incurring high development costs, hence prospects for deploying differentiated (but not necessarily narrow) AI systems with a collectively comprehensive range of task capabilities may best be approached through an explicit consideration of potential AI research and development processes.

15.4 AI development naturally produces structured AI systems

Today, the AI research community is developing a growing range of relatively narrow, strongly differentiated AI systems and composing them to build systems that embrace broader domains of competence. A system that requires, for example, both vision and planning will contain vision and planning components; a system that interprets voice input will contain interacting yet distinct speech recognition and semantic interpretation components. A self-driving car with a conversational interface would include components with all of the above functionalities, and more. The principle of composing differentiated competencies to implement broader task-performance naturally generalizes to potential systems that would perform high-order tasks such as the human-directed design and management of space transportation systems, or AI research and development.

15.5 Structure arises from composing components, not partitioning unitary systems

If one begins with unitary systems as a reference model, the task of implementing structured, broadly competent AI systems may appear to be a problem of *imposing structure by functional decomposition*, rather than one of *building structures by composing functional components.* In other words, taking a unitary-system model as a reference model focuses attention on how a hypothetical system with unitary, universal competence might be divided into parts. While this framing may be useful in conceptual design, it can easily lead to confusion regarding the nature of structured AI system development and products.

15.6 A development-oriented approach to deeply structured systems suggests a broad range of topics for further inquiry

A focus on AI development links AI safety studies to current R&D practice. In particular, the prospective ability to deliver deeply-structured, task-focused AI systems offers rich affordances for the study and potential implementation of safe superintelligent systems.

The joint consideration of structured AI development and products invites inquiry into a range of topics, including:

- Abstract and concrete models of structured AI systems
- Abstract and concrete models of AI R&D automation
- Incremental R&D automation approaching the recursive regime
- Conditions for problematic emergent behaviors in structured systems
- Applications of end-to-end learning in structured systems
- Unitary models as guides to potential risks in composite systems
- Distinct functionalities and deep component integration
- Safety guidelines for structured AI systems development
- Potential incentives to pursue alternative paths
- Potential incentives to violate safety guidelines
- Implications of safety-relevant learning during AI development
- Applications of task-focused AI capabilities to AI safety problems
- Applications of superintelligent-level machine learning to predicting human approval

Further Reading

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 12: AGI agents offer no compelling value*
- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 39: Tiling task-space with AI services can provide general AI capabilities*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*

16 Aggregated experience and centralized learning support AI-agent applications

> 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

Since Turing, discussions of advanced AI have tacitly assumed that agents will learn and act as individuals; naïve scaling to multi-agent systems retains a

human-like model centering on individual experience and learning. The development of self-driving vehicles, however, illustrates a sharply contrasting model in which aggregation of information across N agents (potentially thousands to millions) speeds the acquisition of experience by a factor of N, while centralized, large-scale resources are applied to training, and amortization reduces a range of per-agent costs by a factor of $1/N$. In addition to advantages in speed and amortization, centralized learning enables pre-release testing for routinely encountered errors and ongoing updates in response to rarely-encountered events. The strong, generic advantages of aggregated experience and centralized learning have implications for our understanding of prospective AI-agent applications.

16.2 Discussions often assume learning centered on individual agents

Advanced AI agents are often modeled as individual machines that learn tasks in an environment, perhaps with human supervision, along the lines suggested by Fig. 1. In human experience, human beings have been the sole intelligent agents in the world, a circumstance that powerfully reinforces our habit of identifying experience and learning with individual agents.

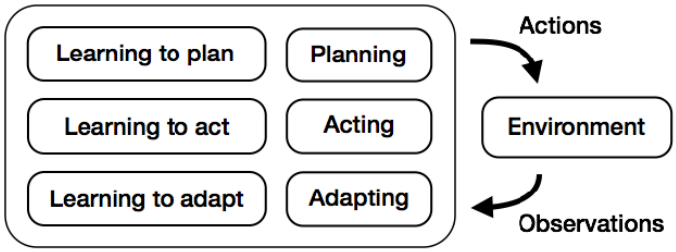

Figure 6: *Individual agents capable of open-ended learning would plan, act, and adapt their actions to a particular task environment, while building on individual experience to learn better methods for planning, acting, and adaptation (generic task learning).*

16.3 Naïve scaling to multi-agent systems replicates individual agents

Naïve scaling of the individual-agent model to multiple agents does not fundamentally alter this picture (Fig 2). One can extend the environment of each agent to include other AI agents while retaining a human-like model of learning: Other agents play a role like that of other human beings.

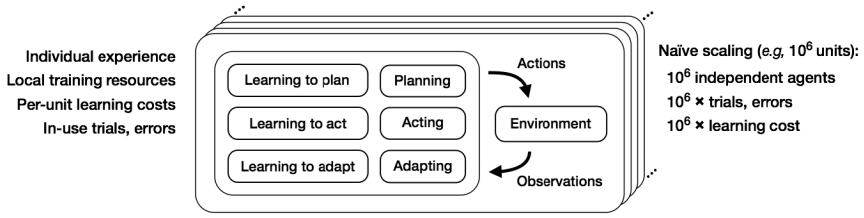

Figure 7: *In a naïve scale-up of the scheme outlined in Fig. 1,* N *agents would plan, act, and adapt their actions to a range of similar task environments while learning from experience independently; both costs and benefits scale as* N*, and the time required to learn tasks remains unchanged.*

16.4 Self-driving vehicles illustrate the power of aggregating experience

Self-driving cars are agents that follow a quite different model: They do not learn as individuals, but instead deliver improved competencies through a centralized R&D process that draws on the operational experience of many vehicles. Tesla today produces cars with self-driving hardware at a rate of approximately 100,000 per year, steadily accelerating the accumulation of driving experience (termed "fleet learning"). Vehicle-agents that learned only from individual experience could not compete in performance or safety.

16.5 Efficiently organized machine learning contrasts sharply with human learning

Machine learning contrasts sharply with human learning in its potential for efficiently aggregating experience, amortizing learning, applying population-based exploration (Conti et al. 2017), and reproducing and distributing competencies.

Experience can be aggregated.
Among human beings, the ability to share detailed experiences is limited, and learning from others competes with learning from experience. In machine learning, by contrast, experience can be aggregated, and learning from aggregated experience need not compete with the acquisition of further experience.

Learning can be accelerated and amortized.
Among human beings, applying a thousand brains to learning does not reduce individual learning time or cost. In machine learn-

ing, by contrast, applying increasing computational capacity can reduce learning time, and computational costs can be amortized across an indefinitely large number of current and future systems.

Competencies can be reproduced.

Among human beings, training each additional individual is costly because competencies cannot be directly reproduced (hence learning elementary mathematics absorbs millions of person-years per year). In machine learning, learned competencies can be reproduced quickly at the cost of a software download.

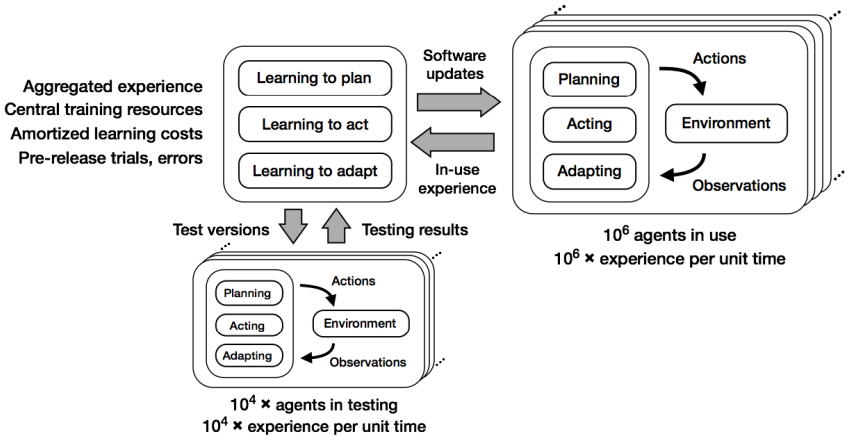

Figure 8: *In large-scale agent applications,* N *agents (e.g. a thousand or a million) would independently plan, act, and adapt in a range of similar task environments, while aggregation of the resulting task experience enables data-rich learning supported by centralized development resources. Centralized learning enables upgraded agent software to be tested before release.*

16.6 Aggregated learning speeds development and amortizes costs

With learning aggregated from N agents, the time required to gain a given quantity of experience scales as $1/N$, potentially accelerating development of competencies. Further, the computation costs of training can be amortized over N agents, yielding a per-agent cost that scales as $1/N$.

If common situations each require human advice when first encountered, the burden of supervising an independent agent might be intolerable, yet acceptably small when employing an agent trained with ongoing experience

aggregated from a large-N deployment. Similarly, if the causes of novel errors can be promptly corrected, the per-agent probability of encountering a given error will be bounded by $1/N$.

16.7 The advantages of aggregated, amortized learning have implications for prospective AI-agent applications

The advantages of richer data sets, faster learning, and amortization of the costs of training and supervision all strongly favor development approaches that employ centralized, aggregated learning across deployments of agents that share similar tasks. These considerations highlight the importance of development-oriented models in understanding prospects for the emergence of broad-spectrum AI-agent applications.

Further Reading

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents*
- *Section 23: AI development systems can support effective human guidance*
- *Section 36: Desiderata and directions for interim AI safety guidelines*

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

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

17.1 Summary

Advances in deep reinforcement learning and end-to-end training have raised questions regarding the likely nature of advanced AI systems. Does progress in deep RL naturally lead to undifferentiated, black-box AI systems with broad capabilities? Several considerations suggest otherwise, that RL techniques will instead provide task-focused competencies to heterogeneous systems. General AI services must by definition encompass broad capabilities, performing not

a single task trained end-to-end, but many tasks that serve many ends and are trained accordingly. Even within relatively narrow tasks, we typically find a range of distinct subtasks that are best learned in depth to provide robust functionality applicable in a wider range of contexts. We can expect to see RL applied to the development of focused systems (whether base-level or managerial) with functionality that reflects the natural diversity and structure of tasks.

17.2 RL and end-to-end training tend to produce black-box systems

Methods that employ end-to-end training and deep reinforcement learning (here termed simply "deep RL") have produced startling advances in areas that range from *game play* (Mnih et al. 2015) to *locomotion* (Heess et al. 2017) to *neural-network design* (Zoph and Le 2016). In deep RL, what are effectively black-box systems learn to perform challenging tasks directly from reward signals, bypassing standard development methods. Advances in deep RL have opened fruitful directions for current research, but also raise questions regarding the likely nature (and safety) of advanced AI systems with more general competencies.

17.3 RL and end-to-end training are powerful, yet bounded in scope

Complex products (both hardware and software) have generally been built of components with differentiated competencies. End-to-end training challenges this model: Although systems are commonly differentiated in some respects (*e.g.*, convolutional networks for visual processing, recurrent neural networks for sequential processing, external memories for long-term representation), these system components and their learned content do not align with distinct tasks at an application level. Functional competencies are (at least from an external perspective) undifferentiated, a confronting us with black-box systems.

Will end-to-end training of black-box systems scale to the development of AI systems with extremely broad capabilities? If so—and if such methods were to be both efficient by metrics of development cost and effective by metrics of product quality—then advanced AI systems might be expected to lack the engineering affordances provided by differentiated systems. In particular, component functionalities might not be identifiable, separable, and subject to intensive training and testing.

There is, however, reason to expect that broad AI systems will comprise patterns of competencies that reflect (and expose[1]) the natural structure of complex tasks.[2] The reasons include both constraints (the nature of training and bounded scope of transfer learning) and opportunities (the greater robustness and generalization capabilities of systems that exploit robust and general components).

17.4 General capabilities comprise many tasks and end-to-end relationships

What would it even mean to apply end-to-end training to a system with truly general capabilities? Consider a hypothetical system intended to perform a comprehensive range of diverse tasks, including conversation, vehicle guidance, AI R&D,[3] and much more. What input information, internal architecture, output modalities, and objective functions would ensure that each task is trained efficiently and well? Given the *challenges of transfer learning* (Teh et al. 2017) even across a range of similar games, why would one expect to find a compelling advantage in learning a comprehensive range of radically different tasks through end-to-end training of a single, undifferentiated system?

Diverse tasks encompass many end-to-end relationships. A general system might provide services that include witty conversation and skillful driving, but it is implausible that these services could best be developed by applying deep RL to a single system. Training a general system to exploit differentiated resources (providing knowledge of vehicle dynamics, scene interpretation, predictions of human road behavior; linked yet distinct resources for conversing with passengers about travel and philosophy) seems more promising than attempting to treat all these tasks as one.

17.5 Broad capabilities are best built by composing well-focused competencies

Systems that draw on (and perhaps adapt) distinct subtask competencies will often support more robust and general performance. For example, to interact with human beings well calls for a model of many aspects of human

1. Section 9: Opaque algorithms are compatible with functional transparency and control
2. Section 38: Broadly-capable systems coordinate narrower systems
3. Section 10: R&D automation dissociates recursive improvement from AI agency

concerns, capabilities, intentions, and responses to situations—aspects that are unlikely to be thoroughly explored through deep RL within the scope of any particular task. For example, a model of an open task environment, such as a road, may fail to model child ball-chasing events that are rare on roads, but common on playgrounds. Similarly, a system intended to explore theoretical physics might struggle to discover mathematical principles that might better be provided through access to a system with strong, specifically mathematical competencies. The use of focused, broadly-applicable competencies in diverse contexts constitutes a powerful form of transfer learning.

Narrow components can support—and strengthen—broad capabilities, and are best learned in depth and with cross-task generality, not within the confines of a particular application. Note that components can be distinct, yet deeply integrated[1] at an algorithmic and representational level.

17.6 Deep RL can contribute to R&D automation within the CAIS model of general AI

Reinforcement learning fits naturally with the R&D automation model of comprehensive AI services.[2] Deep RL has already been applied to *develop state-of-the art neural networks* (Zoph and Le 2016), including *scalable modular systems* (Zoph et al. 2017), and deep RL has been applied to *optimizing deep RL systems*. Increasing automation of AI R&D will facilitate the development of task-oriented systems of all kinds, and will naturally result in deeply-structured systems.[3]

In considering RL in the context of AI control and safety, in is important to keep in mind that RL systems are not utility-maximizing agents,[4] that learning is separable from performance,[5] that human oversight need not impede rapid progress,[6] and that component-level algorithmic opacity is compatible with

1. Section 9: Opaque algorithms are compatible with functional transparency and control
2. Section 12: AGI agents offer no compelling value
3. Section 15: Development-oriented models align with deeply-structured AI systems
4. Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents
5. Section 2: Standard definitions of "superintelligence" conflate learning with competence
6. Section 24: Human oversight need not impede fast, recursive AI technology improvement

system-level functional transparency.[1] Because efficiency, quality, reliability, and safety all favor the development of functionally differentiated AI services, powerful RL techniques are best regarded as tools for implementing and improving AI services, not as harbingers of omnicompetent black-box AI.

Further Reading

- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 9: Opaque algorithms are compatible with functional transparency and control*
- *Section 12: AGI agents offer no compelling value*
- *Section 18: Reinforcement learning systems are not equivalent to reward-seeking agents*
- *Section 38: Broadly-capable systems coordinate narrower systems*

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

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

18.1 Summary

Reward-seeking reinforcement-learning agents can in some instances serve as models of utility-maximizing, self-modifying agents, but in current practice, RL systems are typically distinct from the agents they produce, and do not always employ utility-like RL rewards. In multi-task RL systems, for example, RL "rewards" serve not as sources of value to agents, but as signals that guide training, and unlike utility functions, RL "rewards" in these systems are neither additive nor commensurate. RL systems *per se* are not reward-seekers (instead, they *provide* rewards), but are instead running instances of algorithms that can be seen as evolving in competition with others, with implementations subject to variation and selection by developers. Thus, in current RL practice, developers, RL systems, and agents have distinct purposes and roles.

1. Section 9: Opaque algorithms are compatible with functional transparency and control

18.2 Reinforcement learning systems differ sharply from utility-directed agents

Current AI safety discussions sometimes treat RL systems as agents that seek to maximize reward, and regard RL "reward" as analogous to a utility function. Current RL practice, however, diverges sharply from this model: RL systems comprise often-complex training mechanisms that are fundamentally distinct from the agents they produce, and RL rewards are not equivalent to utility functions.

18.3 RL systems are neither trained agents nor RL-system developers

In current practice, RL systems and task-performing agents often do not behave as unitary "RL agents"; instead, trained agents are products of RL systems, while RL systems are products of a development process. Each of these levels (development processes, RL systems, and task-performing agents) is distinct in its implementation and implicit goals.

18.4 RL systems do not seek RL rewards, and need not produce agents

RL-system actions include running agents in environments, recording results, and running RL algorithms to generate improved agent-controllers. These RL-system actions are not agent-actions, and rewards to agents are not rewards to RL systems. Running agents that collect rewards is a training cost, not a source of reward to the training system itself.

Note that the products of RL systems *need not be agents:* For example, researchers have applied RL systems to train mechanisms for attention in vision networks (Xu et al. 2015), to direct memory access in memory-augmented RNNs (Gülçehre, Chandar, and Bengio 2017), and (in meta-learning) to develop RL algorithms in RL systems (Duan et al. 2016; J. X. Wang et al. 2016). RL systems have also been used to design architectures for convolutional neural networks (Baker et al. 2016; Zoph and Le 2016) and LSTM-like recurrent cells for natural-language tasks (Zoph and Le 2016).

18.5 RL rewards are not, in general, treated as increments in utility

"Reward" must not be confused with utility. In DeepMind's work on multi-task learning, for example, agents are trained to play multiple Atari games

with incommensurate reward-scores. Researchers have found that these heterogeneous "rewards" cannot be scaled and summed over tasks as if they were measures of utility, but must instead be adaptively adjusted to provide learning signals that are effective across different games and stages of training [ref]. RL rewards are sources of information and direction for RL systems, but are not sources of value for agents. Researchers often employ "reward shaping" to direct RL agents toward a goal, but the rewards used shape the agent's behavior are conceptually distinct from the value of achieving the goal.[1]

18.6 Experience aggregation blurs the concept of individual reward

Modern RL systems typically aggregate experience[2] across multiple instances of agents that run in parallel in different environments. An agent-instance does not learn from "its own" experience, and aggregated experience may include off-policy actions that improve learning, yet impair reward-maximization for any given instance.

18.7 RL algorithms implicitly compete for approval

RL algorithms have improved over time, not in response to RL rewards, but through research and development. If we adopt an agent-like perspective, RL algorithms can be viewed as competing in an evolutionary process where success or failure (being retained, modified, discarded, or published) depends on developers' approval (not "reward"), which will consider not only current performance, but also assessed novelty and promise.

18.8 Distinctions between system levels facilitate transparency and control

The patterns and distinctions described above (developer *vs.* learning system *vs.* agent) are not specific to RL, and from a development-oriented perspective, they seem generic. Although we can sometimes benefit from dropping these distinctions and exploring models of agent-like RL systems that seek

1. For example, an RL system can learn a predictive model of a human observer's approval at the level of actions, learning to perform difficult tasks without a specified goal or reward: see Christiano et al. (2017).

2. Section 16: Aggregated experience and centralized learning support AI-agent applications

utility-like rewards, current and future RL systems need not conform to those models. AI development practice suggests that we also consider how AI components and systems are architected, trained, combined, and applied. A development-oriented perspective[1] focuses attention on structured processes, structured architectures, and potential points of control that may prove useful in developing safe applications of advanced AI technologies.

18.9 RL-driven systems remain potentially dangerous

It should go without saying that RL algorithms could serve as engines driving perverse behavior on a large scale. For an unpleasantly realistic example, consider the potential consequences of giving an RL-driven system read/write access to the internet—including access to contemporaneous AI services—with the objective of maximizing the net flow of money into designated accounts. In this scenario, consider how the value of a short position could be increased by manipulating news or crashing a power grid. Distinctions between system levels offer affordances for control, yet levels can be collapsed, and having affordances for control in itself precludes neither accidents nor abuse.

Further Reading

- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 16: Aggregated experience and centralized learning support AI-agent applications*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*
- *Section 36: Desiderata and directions for interim AI safety guidelines*
- *Section 40: Could 1 PFLOP/s systems exceed the basic functional capacity of the human brain?*

19 The orthogonality thesis undercuts the generality of instrumental convergence

> 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.

1. Section 15: Development-oriented models align with deeply-structured AI systems

19.1 Summary

Bostrom (2014) presents carefully qualified arguments regarding the "orthogonality thesis" and "instrumental convergence", but the scope of their implications has sometimes been misconstrued. The orthogonality thesis proposes that any level of intelligence can be applied to any goal (more or less), and the principle of instrumental convergence holds that a wide range of goals can be served by the pursuit of subgoals that include self preservation, cognitive enhancement, and resource acquisition. This range of goals, though wide, is nonetheless limited to goals of indefinite scope and duration. The AI-services model suggests that essentially all practical tasks are (or can be) directly and naturally bounded in scope and duration, while the orthogonality thesis suggests that superintelligent-level capabilities can be applied to such tasks. At a broad, systemic level, tropisms toward general instrumental subgoals seem universal, but such tropisms do not imply that a diffuse system has the characteristics of a willful superintelligent agent.

19.2 *The thesis:* Any level of intelligence can be applied to any goal (more or less)

The orthogonality thesis (Bostrom 2014, p.107) proposes that intelligence and final goals are orthogonal: "[...] more or less any level of intelligence can be combined with more or less any final goal."

A natural consequence of the orthogonality thesis is that intelligence of any level can be applied to goals that correspond to tasks of bounded scope and duration.

19.3 A wide range of goals will engender convergent instrumental subgoals

As Marvin Minsky noted in a conversation *ca.* 1990, a top-level goal of narrow scope (*e.g.*, playing the best possible game of chess) can be served by a subgoal of enormous scope (*e.g.*, converting all accessible resources into chess-playing machinery). The instrumental convergence (IC) thesis (Bostrom 2014, p.109) generalizes this principle; it describes instrumental values that engender subgoals that, if accomplished,

> *[...] would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values [or (sub)goals] are likely to be pursued by a broad spectrum of situated intelligent agents.*

The explicitly considered IC subgoals are:

- ***Self preservation*** (to pursue long-term goals, an agent must continue to exist)
- ***Goal-content integrity*** (to pursue long-term goals, an agent must maintain them)
- ***Cognitive enhancement*** (gaining intelligence would expand an agent's capabilities)
- ***Technological perfection*** (developing better technologies would expand an agent's capabilities)
- ***Resource acquisition*** (controlling more resources would expand an agent's capabilities)

Recognizing the broad scope of IC subgoals provides insight into potential behaviors of a system that pursues goals with a "superintelligent will" (Bostrom 2014, p.105).

19.4 Not all goals engender IC subgoals

As formulated, the IC thesis applies to "a wide range [implicitly, a limited range] of final goals", and the subsequent discussion (Bostrom 2014, p.109) suggests a key condition, that "an agent's final goals concern the future". This condition is significant: Google's neural machine translation system, for example, has no goal beyond translating a given sentence, and the scope of this goal is independent of the level of intelligence that might be applied to achieve it.

In performing tasks of bounded scope and duration, the pursuit of longer-term IC subgoals would offer no net benefit, and indeed, would waste resources. Optimization pressure on task-performing systems can be applied to suppress not only wasteful, off-task actions, but off-task modeling and planning.[1]

19.5 Not all intelligent systems are goal-seeking agents in the relevant sense

As formulated above, the IC thesis applies to "situated agents", yet in many instances intelligent systems that perform tasks are neither agents nor situated in any conventional sense: Consider systems that prove theorems, translate

1. Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects

books, or perform a series of design optimizations. Further, agents within the scope of the IC thesis are typically modeled as rational, utility-directed, and concerned with goals of broad scope, yet even situated agents need not display these characteristics (consider System 1 decision-making in humans).

19.6 Comprehensive services can be implemented by systems with bounded goals

The AI-services model invites a functional analysis of service development and delivery, and that analysis suggests that practical tasks in the CAIS model are readily or naturally bounded in scope and duration. For example, the task of *providing a service* is distinct from the task of *developing a system to provide that service*, and tasks of both kinds must be completed without undue cost or delay. Metalevel tasks such as consulting users to identify application-level tasks and preferences, selecting and configuring systems to provide desired services, supplying necessary resources, monitoring service quality, aggregating data across tasks, and upgrading service-providers are likewise bounded in scope and duration. This brief sketch outlines a structure of generic, bounded tasks that could, by the orthogonality thesis, be implemented by systems that operate at a superintelligent level. It is difficult to identify bounded (*vs.* explicitly world-optimizing) human goals that could more readily be served by other means.

19.7 IC goals naturally arise as tropisms and as intended services

The IC thesis identifies goals that, although though not of value in every context, are still of value at a general, systemic level. The IC goals arise naturally in an AI-services model, not as the result of an agent planning to manipulate world-outcomes in order to optimize an over-arching goal, but as system-level tropisms that emerge from local functional incentives:

- ***Self preservation:*** Typical service-providing systems act in ways that avoid self-destruction: Self-driving cars are an example, though missiles are an exception. AI systems, like other software, can best avoid being scrapped by providing valuable services while not disrupting their own operation.
- ***Goal-content integrity:*** For AI systems, as with other software, functional (implicitly, "goal") integrity is typically critical. Security services that protect integrity are substantially orthogonal to functional services,

however, and security services enable software upgrade and replacement rather than simply preserving what they protect.

- ***Cognitive enhancement:*** At a global level, AI-supported progress in AI technologies can enable the implementation of systems with enhanced levels of intelligence, but most AI R&D tasks are more-or-less orthogonal to application-level tasks, and are bounded in scope and duration.
- ***Technological perfection:*** At a global level, competition drives improvements in both hardware and software technologies; on inspection, one finds that this vast, multi-faceted pursuit resolves into a host of loosely-coupled R&D tasks that are bounded in scope and duration.
- ***Resource acquisition:*** AI systems typically acquire resources by providing value through competitive services (or disservices such as theft or fraud).

All these goals are pursued today by entities in the global economy, a prototypical diffuse intelligent system.

19.8 Systems with tropisms are not equivalent to agents with "will"

The aggregate results of AI-enabled processes in society would tend to advance IC goals even in the absence of distinct AI agents that meet the conditions of the IC thesis. To regard systemic tropisms as manifestations of a "superintelligent will", however, would be much like attributing a "will" to a global ecosystem or economy—a potentially useful perspective that does not reflect an equivalence.

The analogy between *tropisms* and *will* invites a "motte and bailey argument" that wrongly attributes the properties of willful rational agents to *all* systems in which strong aggregate capabilities provide wide-ranging services. Similarly, to argue that a diffuse system would itself undertake actions to *become* a willful agent in order to pursue IC subgoals is essentially circular.

Further Reading

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*
- *Section 12: AGI agents offer no compelling value*

- *Section 28: Automating biomedical R&D does not require defining human welfare*
- *Section 38: Broadly-capable systems coordinate narrower systems*

20 Collusion among superintelligent oracles can readily be avoided

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

The potential for successful collusion among actors decreases as their number increases and as their capabilities, knowledge, situations, and roles become more diverse. In the context of AI systems, these conditions are easy to implement and are often naturally preferred for practical reasons. Accordingly, arguments that sets of high-level AI systems will naturally or inevitably collude are ill-founded, and diverse superintelligent-level question-answering systems ("oracles") can be employed to improve, rather than degrade, safety.

20.2 Trustworthiness can be an emergent property

Prospects for solving AI-safety problems would be greatly improved if we could safely apply SI-level question-answering capabilities to solving those problems. A familiar objection dismisses this potentially powerful approach as unsafe in itself, arguing that, absent solutions to difficult problems, individual SI-level systems would be untrustworthy, and that attempts to establish checks and balances among multiple systems (*e.g.*, through SI-level evaluation of proposed answers to queries) would be thwarted by collusive cooperation.

Identifying robust methods for ensuring non-collusion among SI query-answering systems would overcome this objection, enabling the construction of trustworthy composite oracles and inviting the exploration of SI-enabled strategies for managing potentially untrusted SI systems. The present discussion argues that the *robust non-emergence* of deceptive collusion among SI query-answering systems can be ensured by structuring a set of systems with diverse capabilities, knowledge, situations, and roles.

Given the problem framing, the discussion here will take as non-problematic the ability to develop systems that can respond to queries with

SI-level competence, and (as a consequence of the nature of digital systems) will assume the ability to constrain information inputs to SI systems during their development and use.

In the problematic case, SI query-answering systems might provide deceptive answers in pursuit of emergent, potentially undesirable objectives. The aim here is to identify principles for architecting composite systems that will act as non-deceptive oracles, while treating their components as actors with *potential* incentives for deceptive collusion.

20.3 A range of conditions may facilitate or disrupt collusion

Potential problems of deceptive collusion can be addressed by architecting systems that make collusion difficult or impossible. We can identify a range of relevant conditions:

Conditions that tend to *facilitate* collusion among actors

A1) *Small numbers of actors* (the required scope of collusion is small)
A2) *Insensitivity to defectors* (partial collusion is sufficient)
A3) *Similarity among actors* (*e.g.*, shared histories and objectives)
A4) *Iterated decision processes* with universally observable actions
A5) *Common knowledge* of the above circumstances

Contrasting conditions that tend to *disrupt* collusion among actors

B1) *Large numbers of actors* (the required scope of collusion is large)
B2) *Sensitivity to defectors* (collusion fails unless it is complete)
B3) *Diversity among actors* (*e.g.*, different histories, adversarial objectives)
B4) *Single-move decision processes* in which most actions are unobservable
B5) *Lack of common knowledge* of many strategically-relevant circumstances

The implementation choices implied by conditions (A1–A5) are neither necessary nor desirable, while as suggested by (C1–C5) below, the choices implied by conditions (B1–B5) are both practical and well-motivated.

Natural architectures for problem-solving systems

C1) To improve the quality of answers, it is natural to implement *multiple, diverse* (and implicitly competing) systems to propose alternatives.
C2) To identify low-quality or misleading answers, it is natural to employ diverse critics, *any one of which* could disrupt deceptive collusion.

C3) Systems of diverse, competing proposers and critics naturally implement both *independent* and *adversarial* objectives.

C4) It is natural to apply fixed (hence memory-free) system instantiations to multiple problems, incidentally yielding a series of *history-blind*, *single-move decisions*.

C5) It is natural to provide *differentiated*, *task-relevant information* to systems solving different problems, typically omitting knowledge of general circumstances.

Points (C1–C5) suggest that implementing systems that are resistant to the emergence of deceptive collusion aligns with the general strategy of employing ensembles of diverse, stable, task-directed system components to obtain and assess answers to queries. In appropriate architectures, the use of greater numbers of potentially untrustworthy SI-level components can improve rather than degrade safety.[1]

20.4 Collusion is fragile and easily disrupted

In light of the above considerations, it seems both possible and practical to establish conditions that would preclude deceptive collusion among diverse, task-oriented, SI-level query-answering systems, thereby addressing a key problem in safely applying oracle-class superintelligent capabilities (see Bostrom [2014, p.145]). Prospects for obtaining trustworthy answers from sets of individually untrusted SI systems call for deeper exploration of AI safety strategies premised on reliable access to superintelligent problem-solving capabilities. Capabilities of particular interest include the ability run and evaluate the results of superintelligent-level red-team/blue-team competitions relevant to questions of global strategic stability. Christiano (2015a) considers collusion and a range of related topics; I agree with his gently-stated view that "[...] the implicit argument for the robustness of collusion makes a lot of

1. In a familiar class of worst-case models, systems with general superintelligence would infer extensive knowledge about the world from minimal information, and would choose (if possible) to pursue potentially dangerous goals by manipulating the external environment, *e.g.*, through deceptive answers to queries. In this model, (all?) superintelligent systems, even if almost isolated, would infer the existence of others like themselves, and (all?) would employ a decision theory that induces them to collude (in a coordinated way?) to pursue shared objectives. Even if we grant the initial worst-case assumptions, the argument presented above indicates that systems with these extraordinary capabilities would *correctly* infer the existence of superintelligent-level systems *unlike* themselves (systems with diverse and specialized capabilities, knowledge, and interactions, playing roles that include adversarial judges and competitors), and would correctly recognize that collusive deception is risky or infeasible.

implicit assumptions. If I saw an explicit argument I might be able to assess its explicit assumptions, but for now we don't have one."

Further Reading

- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*
- *Section 12: AGI agents offer no compelling value*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 19: The orthogonality thesis undercuts the generality of instrumental convergence*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*

21 Broad world knowledge can support safe task performance

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

21.1 Summary

Is strongly-bounded world knowledge necessary to ensure strongly-bounded AI behavior? Language translation shows otherwise: Machine translation (MT) systems are trained on general text corpora, and would ideally develop and apply extensive knowledge about the world, yet MT systems perform a well-bounded task, serving as functions of type T :: string → string. Broad knowledge and linguistic competencies can support (rather than undermine) AI safety by enabling systems to learn predictive models of human approval from large corpora.

21.2 Bounding task focus does not require bounding world knowledge

Placing tight bounds on knowledge could be used to restrict AI competencies, and dividing broad AI tasks among components with restricted competencies could help to ensure AI safety (as discussed in Drexler [2015]), yet some

tasks may require broad, integrated knowledge of the human world. General-purpose machine translation (MT) exemplifies a task that calls for integrated knowledge of indefinite scope, but also illustrates another, distinct mechanism for restricting competencies: Task-focused training (also discussed in Drexler [2015]).

21.3 Extensive world knowledge can improve (*e.g.*) translation

Fully-general human-quality machine translation would require understanding diverse domains, which in turn would require knowledge pertaining to human motivations, cultural references, sports, technical subjects, and much more. As MT improves, there will be strong incentives to incorporate broader and deeper world knowledge.

21.4 Current MT systems are trained on open-ended text corpora

Current neural machine translation (NMT) systems gain what is, in effect, knowledge of limited kinds yet indefinite scope through training on large, general text corpora. Google's GNMT architecture has been trained on tens of millions of sentence pairs for experiments and on Google-internal production datasets for on-line application; the resulting trained systems established a new state-of-the-art, approaching the quality of "average human translators" (Wu et al. 2016).

Surprisingly, Google's NMT architecture has been successfully trained, with little modification, to perform bidirectional translation for 12 language pairs (Johnson et al. 2016).

Although the system used no more parameters than the single-pair model (278 million parameters), the multilingual model achieves a performance "reasonably close" to the best single-pair models. Subsequent work (below) developed efficient yet greatly expanded multilingual models that improve on previous single-model performance.

21.5 Current systems develop language-independent representations of meaning

NMT systems encode text into intermediate representations that are decoded into a target language. In Google's multilingual systems, one can compare intermediate representations generated in translating sentences from multiple source languages to multiple targets. Researchers find that, for a given set of equivalent sentences (paired with multiple target languages), encodings

cluster closely in the space of representations, while these clusters are well-separated from similar clusters that represent sets of equivalent sentences with a different meaning. The natural interpretation of this pattern is that sentence-encodings represent meaning in a form that is substantially independent of any particular language.

(It may prove fruitful to train similar NMT models on sets of pairs of equivalent sentences while providing an auxiliary loss function that pushes representations within clusters toward closer alignment. One would expect training methods with this auxiliary objective to produce higher-quality language-independent representations of sentence meaning, potentially providing an improved basis for learning abstract relationships.)

21.6 Scalable MT approaches could potentially exploit extensive world knowledge

NMT systems can represent linguistic knowledge in sets of specialized "experts", and Google's recently described "Sparsely-Gated Mixture-of-Experts" (MoE) approach (Shazeer et al. 2017) employs sets of sets of experts, in effect treating sets of experts as higher-order experts. Human experience suggests that hierarchical organizations of experts could in principle (with suitable architectures and training methods) learn and apply knowledge that extends beyond vocabulary, grammar, and idiom to history, molecular biology, and mathematics.

Notably, Google's MoE system, in which "different experts tend to become highly specialized based on syntax and semantics" has enabled efficient training and application of "outrageously large neural networks" that achieve "greater than 1000× improvements in model capacity [137 billion parameters] with only minor losses in computational efficiency on modern GPU clusters". (There is reason to think that these systems exceed the computational capacity of the human brain.[1])

21.7 Specialized modules can be trained on diverse, overlapping domains

As suggested by the MoE approach in NMT, domain-specialized expertise can be exploited without seeking to establish clear domain boundaries that might support safety mechanisms. For example, it is natural to expect that efficient

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

and effective training will focus on the learning the concepts and language of chemistry (for training some modules), of history (for training others), and of mathematics (for yet others). It is also natural to expect that training modules for expertise in translating chemistry textbooks would benefit from allowing them to exploit models trained on math texts, while performance in the history of chemistry would benefit from access to models trained on chemistry and on general history. Current practice and the structure of prospective task domains[1] suggests that optimal partitioning of training and expertise would be soft, and chosen to improve efficiency and effectiveness, not to restrict the capabilities of any part.

21.8 Safe task focus is compatible with broad, SI-level world knowledge

Machine translation systems today are not agents in any conventional sense of the word, and are products of advances in an AI-development infrastructure, not of open-ended "self improvement" of any distinct system. As we have seen, domain-specific task focus is, in this instance, a robust and natural consequence of optimization and training for a specific task, while high competence in performing that task, employing open-ended knowledge, does not impair the stability and well-bounded scope of the translation task.

It is reasonable to expect that a wide range of other tasks can follow the same basic model,[2] though the range of tasks that would naturally (or could potentially) have this character is an open question. We can expect that broad knowledge will be valuable and (in itself) non-threatening when applied to tasks that range in scope from driving automobiles to engineering urban transportation systems.

Further, broad understanding based on large corpora could contribute to predictive models of human approval[3] that provide rich priors for assessing the desirability (or acceptability) of proposed plans for action by agent, helping to solve a range of problems in aligning AI behaviors with human values. By contrast, similar understanding applied to, for example, unconstrained profit maximization by autonomous corporations, could engender enormous risks.

---

1. Section 15: Development-oriented models align with deeply-structured AI systems
2. Section 12: AGI agents offer no compelling value
3. Section 22: Machine learning can develop predictive models of human approval

21.9 Strong task focus does not require formal task specification

The MT task illustrates how a development-oriented perspective can reframe fundamental questions of task specification. In MT development, we find systems (now significantly automated[1] [Britz et al. 2017]) that develop MT architectures, systems that train those architectures, and (providing services outside R&D labs[2]) the trained-and-deployed MT systems themselves. The nature and scope of the MT task is implicit in the associated training data, objective functions, resource constraints, efficiency metrics, *etc.*, while tasks of the systems that develop MT systems are indirectly implicit in that same MT task (together with metalevel resource constraints, efficiency metrics, *etc.*). Nowhere in this task structure is there a formal specification of what it means to translate a language, or a need to formally circumscribe and limit the task.

Further Reading

- *Section 2: Standard definitions of "superintelligence" conflate learning with competence*
- *Section 7: Training agents in human-like environments can provide useful, bounded services*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 23: AI development systems can support effective human guidance*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*

22 Machine learning can develop predictive models of human approval

> 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

Advanced ML capabilities will precede the development of advanced AI agents, and development of predictive models of human approval need not

1. Section 10: R&D automation dissociates recursive improvement from AI agency
2. Section 24: Human oversight need not impede fast, recursive AI technology improvement

incur agent-related risks. Potential training resources for models of human approval include corpora of text and video that reflect millions of person-years of real and imagined actions, events, and human responses; potential corpora include news, history, fiction, law, philosophy, and more. Limited yet broad predictive models of human (dis)approval could provide commonsense defaults and constraints on both long-term plans and immediate actions. The challenges and potential applications of developing useful models of human approval suggest a range of topics for further consideration and inquiry.

22.2 Advanced ML technologies will precede advanced AI agents

The R&D-automation model of AI development shows how asymptotically-recursive AI-technology improvement could yield superintelligent systems (*e.g.*, machine learning systems) without entailing the use of agents. In this model, agents are potential products, not necessary components.

22.3 Advanced ML can implement broad predictive models of human approval

As Stuart Russell has remarked, AI systems will be able to learn patterns of human approval "Not just by watching, but also by reading. Almost everything ever written down is about people doing things, and other people having opinions about it."

By exploiting evidence from large corpora (and not only text), superintelligent-level machine learning could produce broad, predictive models of human approval and disapproval of actions and events (note that predicting human approval *conditioned on events* is distinct from predicting the events themselves). Such models could help guide and constrain choices made by advanced AI agents, being directly applicable to assessing *intended* consequences of actions.

22.4 Text, video, and crowd-sourced challenges can provide training data

Models of human approval can draw on diverse and extensive resources. Existing corpora of text and video reflect millions of person-years of actions, events, and human responses; news, tweets, history, fiction, science fiction, advice columns, sitcoms, social media, movies, CCTV cameras, legal codes, court records, and works of moral philosophy (and more) offer potential sources of

training data. An interactive crowd-sourcing system could challenge participants to "fool the AI" with difficult cases, eliciting erroneous predictions of approval to enable training on imaginative hypotheticals.

22.5 Predictive models of human approval can improve AI safety

Predictive models of human approval and disapproval could serve as safety-relevant components of structured AI systems. Armstrong's (2017) concept of *low-impact AI systems* points to the potentially robust value of minimizing significant unintended consequences of actions, and models of human approval imply models of what human beings regard as significant. When applied to Christiano's (2014) concept of *approval-directed agents*, general models of human approval could provide *strong priors* for interpreting and generalizing indications of human approval for specific actions in novel domains.[1]

22.6 Prospects for approval modeling suggest topics for further inquiry

The concept of modeling human approval by exploiting large corpora embraces a wide range of potential implementation approaches and applications. The necessary scope and quality of judgment will vary from task to task, as will the difficulty of developing and applying adequate models. In considering paths forward, we should consider a spectrum of prospective technologies that extends from current ML capabilities and training methods to models in which we freely assume superhuman capabilities for comprehension and inference.

In considering high-level capabilities and applications, questions may arise with ties to literatures in psychology, sociology, politics, philosophy, and economics. Models of approval intended for broad application must take account of the diversity of human preferences, and of societal patterns of approval and disapproval of others' preferences.

Criteria for approval may be relatively straightforward for self-driving cars, yet intractable for tasks that might have broad effects on human affairs. For tasks of broad scope, the classic problems of AI value alignment arise, yet some of these problems (*e.g.*, perverse instantiation) could be substantially mitigated by concrete models of what human beings do and do not regard as acceptable.

---

1. Note that learning predictive models of a human observer's approval can enable an RL system to learn difficult tasks: see Christiano et al. (2017).

Further Reading

- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 28: Automating biomedical R&D does not require defining human welfare*

23 AI development systems can support effective human guidance

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

23.1 Summary

Prospects for sharply accelerated development of AI technologies raise questions regarding the effectiveness of human guidance. Basic research requires only loosely coupled human guidance; in AI *application* development, by contrast, human guidance is typically essential to the value and safety of products. When exploring challenging, advanced-AI development scenarios, we should consider potential AI-enabled resources and mechanisms that could help us align AI applications with human goals. A non-exhaustive list includes:

- Strong natural language understanding
- Broad models of human approval and disapproval
- Interactive development of task objectives
- Learning from humans through observation
- Large-scale aggregation of experience and supervision
- Routine but cost-sensitive recourse to human advice
- Adversarial, AI-enabled criticism and monitoring

Prospects for AI-supported human guidance suggest that applications of high-level AI could potentially reduce, rather than increase, the challenges of aligning AI applications with human goals.

23.2 Facilitating human guidance is part of the AI-application development task

Facilitating human guidance a key AI service that can improve tradeoffs between AI development speed and human satisfaction. Models of human guidance in which a person confronts "an AGI" *ab initio* appear neither workable nor realistic; more realistic models can instead consider prospects for diverse AI systems that emerge from structured, AI-enabled development processes in which goal-alignment is part of the development task.

23.3 Development tasks include task selection, system design, training, testing, deployment, in-use feedback, and upgrades

Realistic models of AI application development must consider the pervasive, path-dependent structure of actual system development. Systems are composed of components, and development tasks for both systems and components include function definition and system design, then testing, implementation deployment, in-use feedback, and upgrades (typically performed in iterative, looped, overlapping stages). Prospects for successfully applying AI-enabled capabilities to AI system development are best understood in the context of structured, task-oriented products and development processes.[1]

23.4 We should assume effective use of natural-language understanding

In considering advanced AI capabilities, should assume that a range of current objectives have been achieved, particularly where we see strong progress today. In particular, *natural-language understanding* (Johnson et al. 2016) can support powerful mechanisms for defining AI tasks and providing feedback on AI-system behaviors. Even imperfect language understanding can be powerful because both large text corpora and interactive communication (potentially drawing on the knowledge and expressive capabilities of many individuals) can help to disambiguate meaning.

23.5 Generic models of human (dis)approval can provide useful priors

In some applications, avoiding undesirable AI behaviors will require a broad understanding of human preferences. Explicit rules and direct human instruc-

1. Section 15: Development-oriented models align with deeply-structured AI systems

tion seem inadequate, but (along lines suggested by Stuart Russell [Wolchover 2015]) advanced machine learning could develop broad, predictive models of human approval and disapproval[1] by drawing on large corpora of, for example, news, history, science fiction, law, and philosophy, as well as the cumulative results of imaginative, crowd-sourced challenges. Generic models of human (dis)approval can provide useful priors in defining task-specific objectives, as well as constraints on actions and side-effects.

Note that predicting human approval conditioned on events is distinct from predicting the events themselves. Accordingly, in judging an agent's potential actions, predictive models of approval may fail to reflect unpredictable, unintended consequences, yet be effective in assessing predicted, intended consequences (of, *e.g.*, misguided or perverse plans).

23.6 Bounded task objectives can be described and circumscribed

Describing objectives is an initial step in systems development, and conventional objectives are bounded in terms of purpose, scope of action, allocated resources, and permissible side-effects. Today, software developed for self-driving cars drives particular cars, usually on particular streets; in the future, systems developed to study cancer biology will perform experiments in particular laboratories, systems developed to write up experimental results will produce descriptive text, AI-architecting systems will propose candidate architectures, and so on. Priors based on models of human approval[2] can help AI development systems suggest task-related functionality, while models of human disapproval can can help development systems suggest (or implement) hard and soft constraints; along lines suggested by Armstrong (2013), these can include minimizing unintended effects that people might regard as important.

23.7 Observation can help systems learn to perform human tasks

For task that humans can perform, human behavior can be instructive, not only with respect to means (understanding actions and their effects), but with respect to ends (understanding task-related human objectives). Technical studies of cooperative inverse reinforcement learning (Hadfield-Menell et al. 2016) address problems of learning through task-oriented observation, demonstration, and teaching, while Paul Christiano's work (Christiano 2015b) on scalable

1. Section 22: Machine learning can develop predictive models of human approval
2. Section 22: Machine learning can develop predictive models of human approval

control explores (for example) how observation and human supervision could potentially be extended to challenging AI tasks while economizing on human effort. Generic predictive models of human approval can complement these approaches by providing strong priors on human objectives in performing tasks while avoiding harms.

23.8 Deployment at scale enables aggregated experience and centralized learning

Important AI services will often entail large-scale deployments that enable accelerating learning[1] from instances of success, failure, and human responses (potentially including criticism and advice). In addition to accelerating improvement across large deployments, aggregated experience and learning can increase the benefit-to-cost ratio of using and teaching systems by multiplying the system-wide value of users' correcting and advising individual system-instances while diluting the per-user burdens of encountering correctable errors.

23.9 Recourse to human advice will often be economical and effective

Imperfect models of human approval can be supplemented and improved by recourse to human advice. Imperfect models of approval should contain more reliable models of human concern; an expectation of concern together with uncertainty regarding approval could prompt recourse (without overuse) of human advice. Using advice in learning from aggregated experience[2] would further economize the use of human attention.

23.10 AI-enabled criticism and monitoring can strengthen oversight

Concerns regarding perverse planning by advanced AI agents could potentially be addressed by applying comparable AI capabilities to AI development and supervision. In AI development, the aim would be to understand and avoiding the kinds of goals and mechanisms that could lead to such plans; in

1. Section 16: Aggregated experience and centralized learning support AI-agent applications

2. Section 16: Aggregated experience and centralized learning support AI-agent applications

AI applications, the aim would be to monitor plans and actions and recognize and warn of potential problems (or to intervene and forestall them). This kind of AI-enabled adversarial analysis, testing, monitoring, and correction need not be thwarted by collusion among AI systems,[1] even if these systems operate at superintelligent levels of competence.

23.11 AI-enabled AI development could both accelerate application development and facilitate human guidance

Fast AI technology improvement will increase the scope for bad choices and potentially severe risks. Established practice in system development, however, will favor a measure of intelligent caution, informed by contemporaneous experience and safety-oriented theory and practice.[2] We can expect the temptation to move quickly by accepting risks to be offset to some extent by improved support for goal-aligned function definition, system design, testing, deployment, feedback, and upgrade.

It is very nearly a tautology to observe that the balanced use of powerful AI development capabilities can reduce the cost of producing safe and reliable AI products. Further, the underlying principles appear to scale to AI development technologies that enable the safe implementation of a full spectrum of AI services with superhuman-level performance. This potential, of course, by no means assures the desired outcome.

Further Reading

- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 28: Automating biomedical R&D does not require defining human welfare*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*
- *Section 36: Desiderata and directions for interim AI safety guidelines*

---

1. Section 20: Collusion among superintelligent oracles can readily be avoided
2. Section 35: Predictable aspects of future knowledge can inform AI safety strategies

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

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

It has been suggested that competitive pressures would favor fully automated AI technology development, minimizing human involvement in favor of speed, and potentially sacrificing safety. *Technology* development, however, differs from *application* development. Improvement of core research-enabling AI technologies (*e.g.*, algorithms, architectures, training methods, and development infrastructure) need not be directly linked to applications, hence need not have direct, potentially problematic effects on the world. In considering human involvement in R&D, we can distinguish between *participation*, *guidance*, and *monitoring*. Here, *participation* acts within (and potentially delays) a process, *guidance* sets objectives for a process, and *monitoring* enables, for example, safety-oriented interventions. Both *guidance* and *monitoring* can operate outside core technology-development loops, hence need not impose delays; *participation* is optional, and can be a contribution rather than an impediment. Increasing development speed by relaxing in-the-loop human participation need not sacrifice guidance or safety-oriented monitoring. Automation of *world-oriented application development* presents different challenges and risks.

24.2 Must pressure to accelerate AI technology development increase risk?

Technical and economic objectives will continue to drive incremental yet potentially thorough automation of AI R&D. In considering asymptotically recursive automation of AI R&D,[1] it is natural to think of ongoing human involvement as a source of safety, but also of delays, and to ask whether competitive pressures to maximize speed by minimizing human involvement will incur risks. "AI R&D", however, embraces a range of quite different tasks, and different modes of human involvement differ in their effects on speed and safety. Understanding the situation calls for a closer examination.

1. Section 10: R&D automation dissociates recursive improvement from AI agency

24.3 Basic technology research differs from world-oriented applications

It is important to distinguish *basic AI technology* development from *AI application* development. Recursive technology improvement entails the development of AI technologies that support AI technology development; in this process, basic AI technologies (*e.g.*, algorithms, training methods, ML components, and AI R&D tools) serve as both products and components of recursive R&D loops, while world-oriented application development operates outside those loops, building on their products. Thus, recursive R&D loops need not be directly linked to world-oriented applications, hence need not have direct effects on the world. When we ask whether human involvement might slow recursive improvement, we are asking a question about basic AI technology research; when we ask about risks, we are typically concerned with AI applications.

24.4 We can distinguish between human *participation, guidance,* and *monitoring*

In considering human involvement in R&D, we can distinguish between participation, guidance, and monitoring. Here, *participation* implies human actions within an R&D loop, potentially causing delays; *guidance* means setting objectives and assessing results (*e.g.*, by evaluating the performance of new ML components in applications development) in order to orient research; *monitoring* means observation (*e.g.*, of information flows, resource allocation, and the capabilities of delivered components) in order to prompt safety-oriented interventions.

24.5 Guidance and monitoring can operate outside the central AI R&D loop

Both guidance and monitoring operate outside the R&D loop for basic technology improvement: *Guidance* directs R&D toward valuable outputs, not by direct involvement in R&D loops, but by providing feedback signals that (for example) help to train operational R&D components or help R&D-management components allocate resources toward productive activities. *Monitoring* seeks to forestall potential risks, not by direct involvement in R&D loops, but by enabling interventions that forestall misbehavior mediated by R&D outputs. The nature of potential misbehaviors and requisite monitoring and interventions is outside the scope of the present discussion; policies would presumably be

informed by contemporaneous, cumulative experience and safety research.[1]

24.6 Fast, asymptotically-recursive basic research need not sacrifice safety

Because neither guidance nor monitoring operates inside an R&D loop, human involvement need not entail delays. Thus, if fully recursive technology improvement becomes both feasible and desirable, maximizing the rate of progress in basic AI technologies need not sacrifice potentially safety-critical human roles.

24.7 World-oriented applications bring a different range of concerns

Application development and deployment will typically have direct effects on the human world, and many applications will call for iterative development with extensive human involvement. World-oriented application development operates outside the basic-technology R&D loop, placing it beyond the scope of the present discussion.

Further Reading

- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 23: AI development systems can support effective human guidance*
- *Section 28: Automating biomedical R&D does not require defining human welfare*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*

25 Optimized advice need not be optimized to induce its acceptance

> 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.

1. Section 35: Predictable aspects of future knowledge can inform AI safety strategies

25.1 Summary

To optimize advice to produce a result entails optimizing the advice to ensure its acceptance, and hence to manipulate the clients' choices. Advice can instead be optimized to produce a result *conditioned on the advice being accepted*; because the expected value of an outcome conditioned on an action is independent of the probability of the action, there is then no value in manipulating clients' choices. In an illustrative (and practical) case, a client may request advice on options that offer different trade-offs between expected costs, benefits, and risks; optimization of these options does not entail optimization to manipulate a client's choice among them. Manipulation remains a concern, however: In a competitive situation, the most popular systems may optimize advice for seduction rather than value. Absent effective counter-pressures, competition often will (as it already does) favor the deployment of AI systems that strongly manipulate human choices.

25.2 Background (1): Classic concerns

"Oracles" (Bostrom 2014) are a proposed class of high-level AI systems that would provide answers in response to queries by clients; in the present context, oracles that provide advice on how to achieve goals are of particular interest. It has sometimes been suggested that oracles would be safer than comparable agents that act in the world directly, but because oracles inevitably affect the world through their clients' actions, the oracle/agent distinction *per se* can blur. To clarify this situation, it is important to consider (without claiming novelty of either argument or result) whether optimizing oracles to produce effective advice entails their optimizing advice to affect the world.

25.3 Background (2): Development-oriented models

In the RDA-process model,[1] *research* produces basic components and techniques, *development* produces functional systems, and *application* produces results for users. In AI development, an advisory oracle will be optimized for some purpose by a chain of systems that are each optimized for a purpose:

- ***AI research*** optimizes components and techniques to enable development of diverse, effective AI systems.
- ***Advisory-oracle development*** optimizes systems to suggest options for action across some range of situations and objectives.

1. Section 10: R&D automation dissociates recursive improvement from AI agency

- ***Advisory-oracle application*** suggests actions optimized to achieve given objectives in specific situations.

Each stage in an RDA process yields products (components, oracles, advice) optimized with respect to performance metrics (a.k.a. loss functions).

25.4 Optimization *for results* favors manipulating clients' decisions

Giving advice may itself be an action intended to produce results in the world. To optimize advice to produce a result, however, entails optimizing the advice to ensure that the advice is applied. In current human practice, advice is often intended to manipulate a client's behavior to achieve the advisor's objective, and a superintelligent-level AI advisor could potentially do this very well. At a minimum, an oracle that optimizes advice to produce results can be expected to distort assessments of costs, benefits, and risks to encourage fallible clients to implement supposedly "optimal" policies. A range of standard AI-agent safety problems (*e.g.*, perverse instantiation and pursuit of convergent instrumental goals) then arise with full force.

Optimizing oracles to produce advice intended to produce results seems like a bad idea. We want to produce oracles that are not designed to deceive.

25.5 Optimization for results *conditioned on actions* does not entail optimization to manipulate clients' decisions

Oracles could instead be optimized to offer advice that in turn is optimized, not to produce results, but to produce results contingent on the advice being applied. Because the expected value of an outcome conditioned on an action is independent of the probability of the action, optimal advice will not be optimized to manipulate clients' behavior.

25.6 Oracles can suggest options with projected costs, benefits, and risks

Because human beings have preferences that are not necessarily reducible to known or consistent utility functions, it will be natural to ask advisory-oracles to suggest and explain sets of options that offer different, potentially incommensurate combinations of costs, benefits, and risks; thus, useful advice need not be optimized to maximize a predefined utility function, but can instead be judged by Pareto-optimality criteria. With optimization of outcomes conditioned on acceptance, the quality of assessment of costs, benefits,

and risks will be limited by AI competencies, undistorted by a conflicting objective to manipulate clients' choices among alternatives. (Note that the burden of avoiding options that perversely instantiate objectives rests on the quality of the options and their assessment,[1] not on the avoidance of choice-manipulation.)

25.7 Competitive pressures may nonetheless favor AI systems that produce perversely appealing messages

If multiple AI developers (or development systems) are in competition, and if their success is measured by demand for AI systems' outputs, then the resulting incentives are perverse: Advice that maximizes appeal will often be harmful, just as news stories that maximize attention often are false.

Because AI-enabled communications will permit radical scaling of deception in pursuit of profit and power, it seems likely that *human-driven* applications of these capabilities will be the leading concern as we move forward in AI technology. It seems likely that effective countermeasures will likewise require AI-enabled communication that influences large audiences.

Further Reading

- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 23: AI development systems can support effective human guidance*

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

> 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

Design engineering provides a concrete example of a planning task that could benefit from superintelligent-level support, but interactive systems for

1. Section 22: Machine learning can develop predictive models of human approval

performing such tasks match classic templates for emergent AI agency and risk. Nonetheless, examining design systems at the level of task requirements, component capabilities, and development processes suggests that classic AI-agent risks need not arise. In design engineering, effective human oversight is not an impediment, but a source of value. Superintelligent-level services can help solve (rather than create) AI-control problems; for example, strong models of human concerns and (dis)approval can be exploited to augment direct human oversight.

26.2 Design engineering is a concrete example of a planning task

Planning tasks relate means to ends, and systems-level design engineering offers an illustrative example. Systems engineering tasks are characterized by complex physical and and causal structures that often involve complex and critical interactions with human concerns. As with most planning tasks, system-level design is intended to optimize the application of bounded means (finite materials, time, costs. . .) to achieve ends that are themselves bounded in space, time, and value.

26.3 AI-based design systems match classic templates for emergent AI-agent risk

Classic models of potentially catastrophic AI risk involve the emergence (whether by design or accident) of superintelligent AI systems that pursue goals in the world. Some implementations of question-answering systems ("oracles") could present dangers through their potential ability to recruit human beings to serve as unwitting tools (*Superintelligence*, Bostrom 2014); hazardous characteristics would include powerful capabilities for modeling the external world, formulating plans, and communicating with human beings.

Nonetheless, we can anticipate great demand for AI systems that have all of these characteristics. To be effective, AI-enabled design systems should be able to discuss what we want to build, explore candidate designs, assess their expected performance, and output and explain proposals.

The classic model proposes that these tasks be performed by a system with artificial general intelligence (an AGI agent) that, in response to human requests, will seek to optimize a corresponding utility function over states of the world. Because a fully-general AGI agent could by definition perform absolutely any task with superhuman competence, it requires no further thought to conclude that such an agent could provide engineering design services.

26.4 High-level design tasks comprise distinct non-agent-like subtasks

9 diagrams an abstract, high-level task structure for the development and application of AI-enabled design systems.

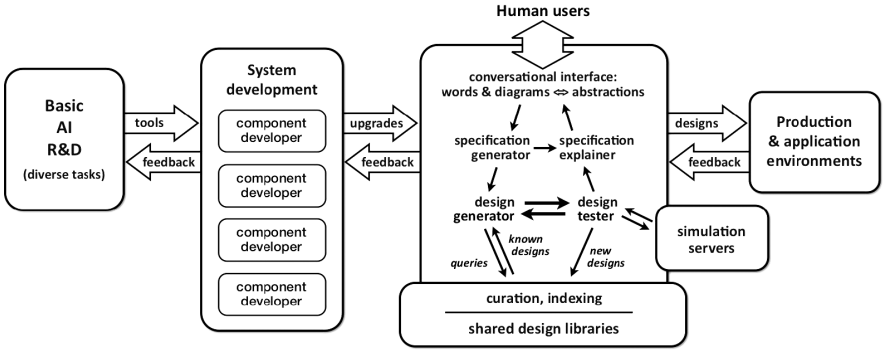

Figure 9: *A task structure & architecture for interactive design engineering*

In this architecture:

- A top-level conversational interface translates between human language (together with sketches, gestures, references to previous designs, *etc.*) and abstract yet informal conceptual descriptions.
- A second level translates between informal conceptual descriptions and formal technical specifications, supporting iterative definition and refinement of objectives, constraints, and general design approach. General AI, *etc.* : Comprehensive AI Services (CAIS)
- The core of the design process operates by iterative generate-and-test, formulating, simulating, and scoring candidate designs with respect to objectives and constraints (including constraints that are tacit and general).
- To enable systems to build on previous results, novel designs can be abstracted, indexed, and cached in a shared library.
- Upstream from design tasks, the development and upgrade of AI-enabled design systems is itself a product of AI-enabled design that integrates progress in basic AI technologies with domain-specific application experience.
- Downstream from design tasks, design products that pass (AI-supported) screening and comparison with competing designs may be deployed and applied, generating application experience that can

inform future design.

In considering the task structure outlined in 9, is important to recognize that humans (today's agents with general intelligence) organize system design tasks in the same general way. The system shown in

9 is *not* more complex than a black-box AGI agent that has acquired engineering competencies; instead, it *makes explicit* the kinds of tasks that must implemented, regardless of how those tasks might be implemented and packaged. Hiding requirements in a black box does not make them go away.

26.5 Real-world task structures favor finer-grained task decomposition

As an aside, we should expect components of engineering systems to be more specialized than those diagrammed above: An any but the most abstract levels, design methods for integrated-circuit design are distinct from methods for aerospace structural engineering, organic synthesis, or AI system architecture, and methods for architecting systems built of diverse subsystems are substantially different from all of these. The degree of integration of components will be a matter of convenience, responsive to considerations that include the value of modularity in fault isolation and functional transparency.

26.6 Use of task-oriented components minimizes or avoids classic AI risks

Consider the flow of optimization and selection pressures implicit in the architecture sketched in 9:

- Systems for basic AI R&D are optimized and selected to produce diverse, high-performance tools (algorithms, generic building blocks. . .) to be used by AI systems that develop AI systems. Off-task activities will incur efficiency costs, and hence will be disfavored by (potentially superintelligent) optimization and selection pressures.
- AI systems that develop AI systems are optimized and selected to produce components and architectures that perform well in applications (here, engineering design). As with basic R&D, off-task activities will incur efficiency costs, and hence will be disfavored.
- All systems, including system-design systems, consist of stable, task-focused components subject to upgrade based on aggregated application experience.

Note that none of these components has a task that includes optimizing either its own structure or a utility function over states of the world.

26.7 Effective human oversight is not an impediment, but a source of value

The role of human oversight it to get what people want, and as such, identifying and satisfying human desires is not an impediment to design, but part of the design process. Comprehensive AI services can include the service of supporting effective human oversight, however, as discussed in safety-oriented requirements for human oversight of basic AI R&D (algorithms, architectures, *etc.*) are minimal,[1] and need not slow progress.

In design, as in other applications of AI technologies, effective human oversight is not enough to avoid enormous problems, because even systems that provide what people think they want can have adverse outcomes. Perversely seductive behaviors could serve the purposes of bad actors, or could arise through competition to develop systems that gain market share (consider the familiar drive to produce news that goes viral regardless of truth, and foods that stimulate appetite regardless of health).

26.8 SI-level systems could solve more AI-control problems than they create

We want intelligent systems that help solve important problems, and should consider how superintelligent-level competencies could be applied to solve problems arising from superintelligence. There is no barrier to using AI to help solve problems of AI control: Deceptive collusion among intelligent problem-solving systems would require peculiar and fragile preconditions.[2] (The popular appeal of "us *vs.* them" framings of AI control is perhaps best understood as a kind of anthropomorphic tribalism.)

26.9 Models of human concerns and (dis)approval can augment direct oversight

In guiding design, key resources will be *language comprehension* and modeling anticipated human (dis)approval.[3] Among the points of potential leverage:

---

1. Section 24: Human oversight need not impede fast, recursive AI technology improvement
2. Section 20: Collusion among superintelligent oracles can readily be avoided
3. Section 22: Machine learning can develop predictive models of human approval

- Strong priors on human concerns to ensure that important considerations are not overlooked.
- Strong priors on human approval to ensure that standard useful features are included by default.
- Strong priors on human disapproval to ensure that options with predictable but excessively negative unintended effects are dropped.
- Building on the above, effective elicitation of human intentions and preferences through interactive questioning and explanation of design options.
- Thorough exploration and reporting (to users and regulators) of potential risks, failure modes, and perverse consequences of a proposal.
- Ongoing monitoring of deployed systems to track unanticipated behaviors, failures, and perverse consequences.

26.10 The pursuit of superintelligent-level AI design services need not entail classic AI-agent risks

To understand alternatives to superintelligent-AGI-agent models, it is best to start with fundamentals—intelligence as problem solving capacity, problems as tasks, AI systems as products of development, and recursive improvement as a process centered on technologies rather than agents. Interactive design tasks provide a natural model of superintelligent-level, real-world problem solving, and within this framework, classic AI-agent problems arise either in bounded contexts, or as a consequence of reckless choices in AI-system development.

Further Reading

- *Section 12: AGI agents offer no compelling value*
- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*

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

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

In a range of competitive, tactical-level tasks (*e.g.*, missile guidance and financial trading), potential advantages in decision speed and quality will tend to favor direct AI control of actions. In high-level strategic decisions, however—where stakes are higher, urgency is reduced, and criteria may be ambiguous—humans can exploit AI competence without ceding control: If AI systems can make excellent decisions, then they can suggest excellent options. Human choice among strategic options need not impede swift response to events, because even long-term strategies (*e.g.*, U.S. nuclear strategy) can include prompt responses of any magnitude. In light of these considerations, we can expect senior human decision makers to choose to retain their authority, and without necessarily sacrificing competitiveness.

27.2 Pressures for speed and quality can favor AI control of decisions

When AI systems outperform humans in making decisions (weighing both speed and quality), competitive situations will drive humans to implement AI-controlled decision processes. The benefits of speed and quality will differ in different applications, however, as will the costs of error.

27.3 Speed is often critical in selecting and executing "tactical" actions

Decision-speed can be critical: Computational systems outperform humans in response time when controlling vehicles and launching defensive missiles, offering crucial advantages, and military planners foresee increasing pressures use AI-directed systems in high-tempo tactical exchanges. In tactical military situations, as in high-frequency financial trading, failure to exploit the advantages of AI control may lead to losses.

27.4 Quality is more important than speed in strategic planning

High-level strategic plans almost by definition guide actions extended in time. Even long-term strategies may of course be punctuated by prompt, conditional actions with large-scale consequences: U.S. nuclear strategy, for example, has been planned and revised over years and decades, yet contemplates swift and overwhelming nuclear counterstrikes. Fast response to events is compatible with deliberation in choosing and updating strategies.

27.5 System that can make excellent decisions could suggest excellent options

Superior reasoning and information integration may enable AI systems to identify strategic options with superhuman speed and quality, yet this need not translate into a pressure for humans to cede strategic control. If AI systems can make excellent decisions, then they can suggest excellent sets of options for consideration by human decision makers.

Note that developing sets of options does not imply a commitment to a utility function over world states; given uncertainties regarding human preferences, it is more appropriate to apply Pareto criteria to potentially incommensurate costs, benefits, and risks, and to offer proposals that need not be optimized to induce particular choices.[1] In this connection, it is also important to remember that long-term strategies are normally subject to ongoing revision in light of changing circumstances and preferences, and hence adopting a long-term strategy need not entail long-term commitments.

27.6 Human choice among strategies does not preclude swift response to change

Profoundly surprising events that call for superhumanly-swift, large-scale, unanticipated strategic reconsideration and response seem likely to be rare, particularly in hypothetical futures in which decision makers have made effective use of superintelligent-quality strategic advice. Further, human choice among strategic options need not be slow: Under time pressure, a decision maker could scan a menu of presumably excellent options and make a quick, gut choice. Beyond this, human-approved strategies could explicitly allow for great flexibility under extraordinary circumstances. In light of these considerations, substantial incentives for routine relinquishment of high-level strategic control seem unlikely.

1. Section 25: Optimized advice need not be optimized to induce its acceptance

27.7 Senior human decision makers will likely choose to retain their authority

Absent high confidence that AI systems will consistently make choices aligned with human preferences, ceding human control of increasingly high-level decisions would incur increasing risks for declining (and ultimately slight) benefits. As a practical matter, we can expect senior human decision makers to choose to retain their authority unless forcibly dislodged, perhaps because they have lost the trust of other, more powerful human decision makers.

Further Reading

- *Section 12: AGI agents offer no compelling value*
- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 23: AI development systems can support effective human guidance*
- *Section 25: Optimized advice need not be optimized to induce its acceptance*

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

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

28.1 Summary

It has been suggested that assigning AGI agents tasks as broad as biomedical research (*e.g.*, "curing cancer") would pose difficult problems of AI control and value alignment, yet a concrete, development-oriented perspective suggests that problems of *general* value alignment can be avoided. In a natural path forward, diverse AI systems would automate and coordinate diverse biomedical research tasks, while human oversight would be augmented by AI tools, including predictive models of human approval. Because strong task alignment does not require formal task specification, a range of difficult problems need not arise. In light of alternative approaches to providing general AI services, there are no obvious advantages to employing risky, general-purpose AGI agents to perform even extraordinarily broad tasks.

28.2 Broad, unitary tasks could present broad problems of value alignment

"The Value Learning Problem" (Soares 2018) opens with an example of a potential AGI problem:[1]

> *Consider a superintelligent system, in the sense of Bostrom (2014), tasked with curing cancer [. . .] without causing harm to the human (no easy task to specify in its own right). The resulting behavior may be quite unsatisfactory. Among the behaviors not ruled out by this goal specification are stealing resources, proliferating robotic laboratories at the expense of the biosphere, and kidnapping human test subjects.*

The following sections will consider biomedical research (including cancer research tasks) from a general but more concrete, less unitary perspective, concluding that undertaking AI-driven biomedical research need not risk programs based on criminality (kidnapping, *etc.*) or catastrophic problems of value alignment. (I thank Shahar Avin for suggesting this topic as a case study.)

Diverse roles and tasks in biomedical research and applications:	
Scientific research:	**Research direction:**
Developing techniques	Resource allocation
Implementing experiments	Project management
Modeling biological systems	Competitors, reviewers
Clinical practice:	**Oversight:**
Patients	Citizen's groups
Physicians	Regulatory agencies
Health service providers	Legislatures

28.3 Diverse AI systems could automate and coordinate diverse research tasks

Biomedical research and applications comprise extraordinarily diverse activities. Development of even a single diagnostic or therapeutic technology, for

1. This quote comes from an earlier draft (*MIRI Technical report 2015–4*), available at `https://intelligence.org/files/obsolete/ValueLearningProblem.pdf`

example, typically draws on many distinct areas of competence that support concept development, experimentation, and data analysis. At a higher level of research organization, project management and resource allocation call for assessment of competing proposals[1] in light of not only their technical merits, but of their costs and benefits to human beings.

Progress in implementing AI systems that provide diverse superhuman competencies could enable automation of a full range of technical and managerial tasks, and because AI R&D is itself subject to automation,[2] progress could be incremental, yet swift. By contrast, it is difficult to envision a development path in which AI developers would treat all aspects of biomedical research (or even cancer research) as a single task to be learned and implemented by a generic system. Prospects for radical improvements in physical tools (*e.g.*, through molecular systems engineering) do not change this general picture.

28.4 Human oversight can be supported by AI tools

Within the scope of biomedical research, several areas—resource investments, *in-vivo* experimentation, and clinical applications—call for strong human oversight, and oversight is typically mandated not only by ethical concerns, but by law, regulation, and institutional rules. Human oversight is not optional (it is part of the task), yet AI applications could potentially make human oversight more effective. For example, systems that describe research proposals[3] in terms of their anticipated human consequences would enable human oversight of complex research plans, while robust predictive models of human concerns[4] could be applied to focus scarce human attention. Consultation with superintelligent-level advisors could presumable enable extraordinarily well-informed judgments by patients and physicians.

28.5 Strong task alignment does not require formal task specification

The example of language translation shows that task alignment need not require formal task specification,[5] and a development-oriented perspective on concrete biomedical tasks suggests that this property may generalize quite

1. Section 12: AGI agents offer no compelling value
2. Section 10: R&D automation dissociates recursive improvement from AI agency
3. Section 23: AI development systems can support effective human guidance
4. Section 22: Machine learning can develop predictive models of human approval
5. Section 21: Broad world knowledge can support safe task performance

widely. Although it is easy to envision both safe and unsafe configurations of task-performing systems, it is reasonable to expect that ongoing AI safety research (both theoretical and informed by ongoing experience) can enable thorough automation of research while avoiding both unacceptable costs and extraordinary risks stemming from emergent behaviors. If safety need not greatly impede development, then unsafe development is best viewed as a bad-actor risk.

28.6 The advantages of assigning broad, unitary tasks to AGI agents are questionable

It has been persuasively argued (Bostrom 2014) that self-improving, general-purpose AGI agents cannot safely be tasked with broad goals, or with seemingly narrow goals that might motivate catastrophic actions. If a full range of superintelligent-level AI capabilities can be provided efficiently by other means,[1] then the advantages of developing risky AGI agents are questionable.

The argument that access to general, superintelligent-level AI capabilities need not incur the risks of superintelligent AGI agents includes the following points:

- Recursive technology improvement[2] is a natural extension of current AI R&D, and does not entail recursive self improvement of distinct AI systems: Agents are products, not development tools.
- Effective human oversight[3] need not substantially impede recursive improvement of *basic AI technologies*, while overseeing the development of *task-focused AI systems* is similar to (but less risky than) specifying tasks for an AGI system.
- Models of human approval[4] can inform AI plans in bounded domains, while the use of AI systems to examine the scope and effects of proposed plans (in an implicitly adversarial architecture[5]) scales to superintelligent proposers and critics.

It seems that there are no clear technical advantages to pursuing an AGI-agent approach to biomedical research, while a task-focused approach provides a

1. Section 12: AGI agents offer no compelling value
2. Section 10: R&D automation dissociates recursive improvement from AI agency
3. Section 24: Human oversight need not impede fast, recursive AI technology improvement
4. Section 22: Machine learning can develop predictive models of human approval
5. Section 20: Collusion among superintelligent oracles can readily be avoided

natural path in which value-alignment problems are bounded and mitigated. Until general value-alignment problems are solved, it would be wise to avoid them.

Further Reading

- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 12: AGI agents offer no compelling value*
- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*

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

> 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

Discussions of SI-level AI technologies and risks have centered on scenarios in which humans confront AGI agents in a world that lacks other, more tractable SI-level AI resources. There is, however, good reason to expect that humans will (or readily could) have access to comprehensive, open-ended AI services before AGI-agent systems are implemented. The open-ended AI-services model of artificial general intelligence does not preclude (and in fact would facilitate) the implementation of AGI agents, but suggests that AI risks, and their intersection with the ethics of computational persons, should be reexamined in the context of an AI milieu that can provide SI-level strategic advice and security services.

29.2 It has been common to envision AGI agents in a weak-AI context

In classic AGI-risk scenarios, advanced AI capabilities emerge in the form of AGI agents that undergo recursive, transformative self-improvement to a

superintelligent (SI) level; these agents then gain capabilities beyond those of both human beings and human civilization. Studies and discussions in this conceptual framework propose that, when confronted with AGI agents, humans would lack prior access to tractable SI-level problem-solving capabilities.

29.3 Broad, SI-level services will (or readily could) precede SI-level AI agents

The technologies required to implement or approximate recursive AI technology improvement are likely to emerge through heterogeneous AI-facilitated R&D mechanisms,[1] rather than being packaged inside a discrete entity or agent. Accordingly, capabilities that could in principle be applied to implement SI-level AGI agents could instead[2] be applied to implement general, comprehensive AI services[3] (CAIS), including stable, task-focused agents.[4] In this model, directly-applied AI services are distinct from services that develop AI services,[5] an approach that reflects natural task structures[6] and has great practical advantages.[7]

29.4 SI-level services will enable the implementation of AGI agents

Although recursive technology improvement will most readily be developed by means of heterogeneous, non-agent systems, any AI milieu that supports "comprehensive AI services" could (absent imposed constraints) provide the service of implementing SI-level AGI agents. This prospect diverges from classic AGI-agent risk scenarios, however, in that a strong, pre-existing AI milieu could be applied to implement SI-level advisory and security services.

---

1. Section 10: R&D automation dissociates recursive improvement from AI agency
2. Section 11: Potential AGI-enabling technologies also enable comprehensive AI services
3. Section 12: AGI agents offer no compelling value
4. Section 16: Aggregated experience and centralized learning support AI-agent applications
5. Section 26: Superintelligent-level systems can safely provide design and planning services
6. Section 38: Broadly-capable systems coordinate narrower systems
7. Section 16: Aggregated experience and centralized learning support AI-agent applications

29.5 SI-level advisory and security services could limit AGI-agent risks

The prospect of access SI-level advisory and security services fundamentally changes the strategic landscape around classic AI-safety problems. For example, "superpowers" as defined by Bostrom (2014) do not exist in a world in which agents lack radically-asymmetric capabilities. Further, arguments that SI-level agents would collude (and potentially provide collectively deceptive advice), do not carry over to SI-level systems in a CAIS milieu,[1] advice can be objective, rather than manipulative,[2] predictive models of human preferences and concerns[3] can improve the alignment of actions with human intentions in performing well-bounded tasks.[4]

Arguments that competitive and security pressures would call for ceding strategic control to AI systems are surprisingly weak: Tactical situations may call for responses of SI-level quality and speed, but SI-level advisory and security services could support strategic choices among excellent options, deliberated at a human pace.[5] Crucially, in a range of potential defense/offense scenarios, the requisite security systems could be endowed with arbitrarily large advantages in resources for strategic analysis, tactical planning, intelligence collection, effector deployment, and actions taken to preclude or respond to potential threats. In choosing among options in the security domain, humans would likely prefer systems that are both reliable and unobtrusive.

Many AI safety strategies have been examined to date, and all have difficulties; it would be useful to explore ways in which tractable SI-level problem-solving capabilities could be applied to address those difficulties. In exploring potential responses to future threats, it is appropriate to consider potential applications of future capabilities.

---

1. Section 20: Collusion among superintelligent oracles can readily be avoided
2. Section 25: Optimized advice need not be optimized to induce its acceptance
3. Section 22: Machine learning can develop predictive models of human approval
4. Section 16: Aggregated experience and centralized learning support AI-agent applications
5. Section 27: Competitive pressures provide little incentive to transfer strategic decisions to AI systems

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

The spectrum of potential AGI systems includes agents that should, from a moral perspective be regarded as persons and treated accordingly. Indeed, the spectrum of potential computational persons includes emulations of generic or specific human beings[1]. To fail to treat such entities as persons would, *at the very least*, incur risks of inadvertently committing grave harm.

It has sometimes been suggested that security in a world with SI-level AI would require stunting, "enslaving", or precluding the existence of computational persons. The prospect of robust, SI-level security services, however, suggests that conventional and computational persons could coexist within a framework stabilized by the enforcement of effective yet minimally-restrictive law.

Bostrom's (2014, p.201–208) concept of "mind crime" presents what are perhaps the most difficult moral questions raised by the prospect of computational persons. In this connection, SI-level assistance may be essential not only to prevent, but to understand the very nature and scope of potential harms to persons unlike ourselves. Fortunately, there is seemingly great scope for employing SI-level capabilities while avoiding potential mindcrime, because computational systems that provide high-order problem-solving services need not be equivalent to minds.[2]

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

The prospect of access to tractable, SI-level capabilities reframes the strategic landscape around the emergence of advanced AI. In this connection, it will be important to reexamine classic problems of AI safety and strategy, not only in the context of an eventual SI-level AI services milieu, but along potential paths forward from today's AI technologies.

Further Reading

- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*

1. For example, "Ems" (Hanson 2016)
2. Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame

- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*
- *Section 12: AGI agents offer no compelling value*
- *Section 16: Aggregated experience and centralized learning support AI-agent applications*
- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 25: Optimized advice need not be optimized to induce its acceptance*
- *Section 26: Superintelligent-level systems can safely provide design and planning services*
- *Section 27: Competitive pressures provide little incentive to transfer strategic decisions to AI systems*
- *Section 38: Broadly-capable systems coordinate narrower systems*

30 Risky AI can help develop safe AI

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

In a familiar AGI threat model, opaque, self-improving AI systems give rise to systems that incorporate wide-ranging information, learn unknown objectives, and could potentially plan to pursue dangerous goals. The R&D-automation/AI-services model suggests that technologies that could enable such systems would first be harnessed to more prosaic development paths, but what if powerful AI-development capabilities were deeply entangled with opaque, untrusted systems? In this event, conceptually straightforward methods could be employed to harness untrusted systems to the implementation of general learning systems that lack problematic information and purposes. Key affordances include re-running development from early checkpoints, and applying optimization pressure with competition among diverse systems to produce compact "learning kernels". Thus, from a path that could lead to problematic AGI agents, a readily accessible off-ramp leads instead to general intelligence in the form of capabilities that enable open-ended AI-service development. *(These and related topics have been explored in an earlier form, but in greater depth, in FHI Technical Report 2015-3 (Drexler 2015).)*

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

In a familiar AGI threat model, the pursuit of general, superintelligent level AI leads to opaque, self-improving systems with ill-characterized information content, inference capabilities, planning capabilities, and goals. Strong arguments suggest that the use of such systems could pose grave risks (Bostrom 2014).

The present analysis will consider a near-worst-case scenario in which AI development stands at the threshold of building such systems, rather than considering how such a situation could be avoided. Further, it will consider the hard case in which "general intelligence" is an indivisible property, a capacity to learn more-or-less anything that a human can, and potentially much more. Crucially, the analysis will neither equate *intelligence as learning capacity* with *intelligence as competence*[1] nor assume that a *product system* must inherit the full information content of the *producing system*. Finally, it will assume that researchers retain copies of precursors of systems of interest.

30.3 Open-ended "self-improvement" implies strong, general AI implementation capabilities

"Self improvement" implies strong, general AI implementation capabilities, yet in considering a chain of improved implementations, the concept of "self" is at best ambiguous, and at worst is an anthropomorphic distraction.[2] Operationally, "self" improvement implies an opaque system that is capable of implementing systems that are better than itself by some metric, and in particular, is capable of implementing systems that are improved in the sense of being better at implementing improved systems (*etc.*). In the proposed AGI threat model, some earlier, non-problematic system was capable of serving as a link in such a chain (given suitable machine resources, training data, simulated environments, and tasks) and the actual development history from that point led to a problematic result.

---

1. Section 2: Standard definitions of "superintelligence" conflate learning with competence

2. Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame

30.4 Successful system development can be recapitulated with variations

In general, results in machine learning can be repeated with variations in architectures, training data, simulated environments, training objectives, *etc.* In the present instance, systems with the potential to give rise to a sequence of improved systems are assumed to be architecturally opaque; nonetheless, external affordances (training data, simulated environments, training objectives, resource-optimization pressures, *etc.*) remain available. Development of powerful systems can be recapitulated with variations induced by external affordances, and these affordances can strongly affect the content of what is learned. If the desired core functionality is development of learning systems, it is likely that relatively abstract problem spaces will sufficient or optimal. In addition, learning to optimize systems for learning a task does not require access to detailed task-level training data or environments, and optimizing systems for the task of *optimizing architectures for systems that learn a task* is even more abstract and remote from object-level training information. Note that tasks at all levels are by nature bounded with respect to time and resources, and hence do not naturally engender convergent instrumental goals.[1]

In research and development, different versions of systems are always in implicit competition with one another to maximize performance on some bounded task: Those that perform poorly by relevant metrics will be set aside, while versions that perform well will be used to produce (or serve as prototypes for) next-generation systems. Thus, the convergent goal of systems under development is competition to perform bounded tasks, and by the orthogonality thesis (Bostrom 2014), the pursuit of bounded goals can employ arbitrarily high intelligence. In aggregate, such systems will (or readily could) satisfy conditions that exclude collusion.[2]

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

The scheme outlined here centers on competitive optimization of "learning kernels" for "compactness", where a learning kernel is a system that, in conjunction with computational resources, auxiliary components, and a set of benchmark "demo tasks", can produce an expanded set of systems that

1. Section 19: The orthogonality thesis undercuts the generality of instrumental convergence
2. Section 20: Collusion among superintelligent oracles can readily be avoided

perform those tasks at some given level of performance. Here, optimizing for compactness implies minimizing a weighted sum of the kernel's description length (operationally, the number of bits in a string that decompresses and compiles to build the system) and metrics corresponding to the resources it consumes in expansion; in the present context, compactness is assumed to give heavy weight to description length, penalizing resource consumption only to exclude undue computational costs. If kernel expansion can produce a range of systems that perform each of a sufficiently broad range of challenging demo tasks, then the learning kernel can be considered general. Examples of demo-task domains might include:

- Language translation
- Visual perception
- Robotic control
- Spacecraft design
- Chemical synthesis
- Strategic planning
- Theorem proving
- Software development

Note that current practice in AI research does in fact iteratively develop, train, and improve relatively compact learning kernels (architectures and algorithms) that, through training, expand their information content to produce systems that perform tasks in particular domains.

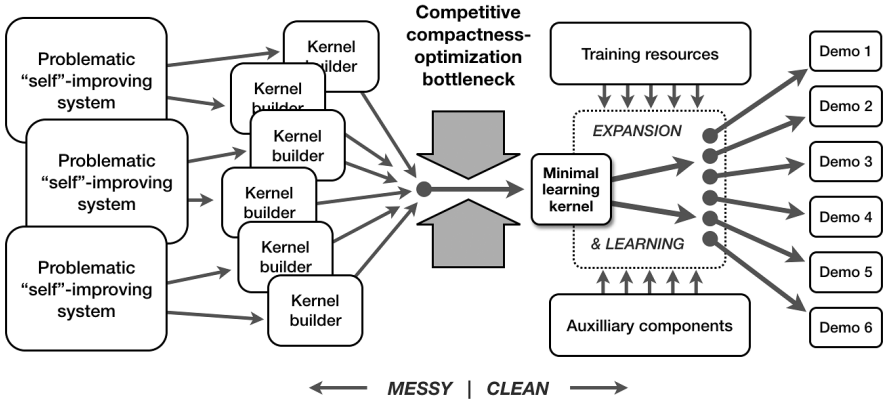

Figure 10: *Applying problematic, ill-characterized AI systems to implementation tasks under competitive optimization pressure could produce clean, compact systems with general learning capabilities. (Schematic diagram)*

30.6 Competitive optimization for compactness can exclude problematic information and competencies

Competition for learning-kernel compactness, subject to retaining generality, will strongly exclude potentially-problematic information that does not contribute to the process of learning target competencies. A world-beating Go system need not "know" that it is "playing a game"; a high-quality machine translation system need not "know" that its outputs are "requested by human beings"; a machine vision system need not "know" that its capabilities enable "autonomous vehicle control". A compact, general kernel that can generate Go-learning, language-learning, and vision-learning systems can (and under optimization, must) "know" even less about concrete tasks, domains, and intended applications in a "human world".

In the standard AGI threat model, self-improvement (which is to say, AI implementation capability) is capable of producing systems that operate at a superintelligent level. Systems at this level that are optimized to optimize general learning kernels for compactness presumably can be quite effective at stripping out information that is unnecessary for task performance. Given that post-bottleneck training data can provide all domain-relevant information, together with auxiliary resources such as efficient numerical algorithms, effective optimization will strip out virtually all world knowledge (geography, history, vocabulary, chemistry, biology, physics. . .), including any plausible basis for problematic plans and concrete world-oriented competencies. This conclusion holds even if the kernel-implementation systems might be untrustworthy if directly applied to world-oriented tasks.

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

In the familiar AGI threat model, development results in opaque, self-improving, ill-characterized, but highly-capable systems, and—crucially—the use of these capabilities is assumed to require that *the problematic systems themselves* be applied to a wide range of world tasks. This assumption is incorrect. As argued above, posited self-improvement capabilities could instead be re-developed from earlier, non-problematic systems through a process that leads to diverse systems that compete to perform tasks that include AI system development. The AI-implementation capacity of such systems could then be applied to the development of compact general-learning kernels that will omit representations of problematic knowledge and goals. This strategy is technology-agnostic: It is compatible

with neural, symbolic, and mixed systems, whether classical or quantum mechanical; it assumes complete implementation opacity, and relies only on optimization pressures directed by external affordances.

Because expansion of a general learning kernel would in effect implement the "research" end of AI R&D automation, the strategy outlined above could provide a clean basis for any of a range of development objectives, whether in an AGI-agent or CAIS model of general intelligence. This approach is intended to address a particular class of scenarios in which development has led the edge of a cliff, and is offered as an example, not as a prescription: Many variations would lead to similar results, and ongoing application of the underlying principles would avoid classic threat models from the start. These safety-relevant principles are closely aligned with current practice in AI system development.[1]

Further Reading

- *Section 1: R&D automation provides the most direct path to an intelligence explosion*
- *Section 5: Rational-agent models place intelligence in an implicitly anthropomorphic frame*
- *Section 8: Strong optimization can strongly constrain AI capabilities, behavior, and effects*
- *Section 9: Opaque algorithms are compatible with functional transparency and control*
- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 11: Potential AGI-enabling technologies also enable comprehensive AI services*
- *Section 36: Desiderata and directions for interim AI safety guidelines*

1. Section 36: Desiderata and directions for interim AI safety guidelines

31 Supercapabilities do not entail "superpowers"

> 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

Superintelligence (Bostrom 2014) develops the concept of "cognitive superpowers" that potentially include intelligence amplification, economic production, technology development, strategic planning, software hacking, or social manipulation. These "superpowers", however, are defined in terms of potential strategic advantage, such that "at most one agent can possess a particular superpower at any given time". Accordingly, in discussing AI strategy, we must take care not to confuse situation-dependent super*powers* with technology-dependent super*capabilities*.

31.2 AI-enabled capabilities could provide decisive strategic advantages

Application of AI capabilities to AI R&D[1] could potentially enable swift intelligence amplification and open a large capability gap between first- and second-place contenders in an AI development race. Strong, asymmetric capabilities in strategically critical tasks (economic production, technology development, strategic planning, software hacking, or social manipulation) could then provide decisive strategic advantages in shaping world outcomes (Bostrom 2014, p.91–104).

31.3 Super*powers* must not be confused with super*capabilities*

Bostrom (2014, p.93) introduces the concept of a "superpower" as a property of "a system that sufficiently excels" in one of the strategically critical tasks, stating that "[a] full-blown superintelligence would greatly excel at all of these tasks," and later explains that "excels" must be understood in a relative sense that entails a strong situational asymmetry (Bostrom 2014, p.104):

1. Section 10: R&D automation dissociates recursive improvement from AI agency

> *[. . .] superpowers [. . .] are possessed by an agent as superpowers only if the agent's capabilities in these areas substantially exceed the combined capabilities of the rest of the global civilization [hence] at most one agent can possess a particular superpower at any given time.*

To avoid confusion, it is important to distinguish between strategically relevant capabilities far beyond those of contemporaneous, potentially superintelligent competitors ("superpowers"), and capabilities that are (merely) enormous by present standards ("supercapabilities"). Supercapabilities are robust consequences of superintelligence, while superpowers—as defined—are consequences of supercapabilities in conjunction with a situation that may or may not arise: strategic dominance enabled by strongly asymmetric capabilities. In discussing AI strategy, we must take care not to confuse prospective technological capabilities with outcomes that are path-dependent and potentially subject to choice.

Further Reading

- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 32: Unaligned superintelligent agents need not threaten world stability*

32 Unaligned superintelligent agents need not threaten world stability

> 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

It is often taken for granted that unaligned superintelligent-level agents could amass great power and dominate the world by physical means, not necessarily to human advantage. Several considerations suggest that, with suitable preparation, this outcome could be avoided:

- Powerful SI-level capabilities can precede AGI agents.
- SI-level capabilities could be applied to strengthen defensive stability.
- Unopposed preparation enables strong defensive capabilities.
- Strong defensive capabilities can constrain problematic agents.

In other words, applying SI-level capabilities to ensure strategic stability could enable us to coexist with SI-level agents that do not share our values. The present analysis outlines general prospects for an AI-stable world, but necessarily raises more questions than it can explore.

32.2 General, SI-level capabilities can precede AGI agents

As has been argued elsewhere, the R&D-automation/AI-services model of recursive improvement and AI applications challenges the assumption[1] that the pursuit of general, SI-level AI capabilities naturally or necessarily leads to classic AGI agents. Today, we see increasingly automated AI R&D applied to the development of AI services, and this pattern will (or readily could) scale to comprehensive, SI-level AI services that include the service of developing new services. By the *orthogonality thesis* (Bostrom 2014), high-level AI services could be applied to more-or-less any range of tasks.

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

World order today—from neighborhood safety to the national security—is imperfectly implemented through a range of defensive services, *e.g.*, local surveillance, self-defense, and police; military intelligence, arms control, and defensive weapon systems. A leading strategic problem today is the offensive potential of nominally defensive systems (deterrence, for example, relies on offensive weapons), engendering the classic "security dilemma" and consequent arms-race dynamics.

Bracketing thorny, path-dependent questions of human perceptions, preferences, objectives, opportunities, and actions, one can envision a state of the world in which SI-level competencies have been applied to implement impartial, *genuinely defensive* security services. Desirable implementation steps and systemic characteristics would include:

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

1. Section 12: AGI agents offer no compelling value

In other words, one can envision an AI-stable world in which well-prepared, SI-level systems are applied to implement services that ensure physical security regardless of the preferences of unaligned or hostile actors. (Note that this does not presuppose a solution to AGI alignment: AI-supported design and implementation[1] of policies for security services[2] need not be equivalent to utility maximization by an AGI agent.[3])

32.4 Unopposed preparation enables strong defensive capabilities

A background assumption in this discussion is that, given access to SI-level capabilities, *potentially enormous* resources (indeed, literally astronomical) could be mobilized to achieve critical civilizational goals that include AGI-compatible strategic stability. In other words, we can expect civilizations as a whole to pursue *convergent instrumental goals* (Bostrom 2014, p.109), and to apply the resulting capabilities.

In this connection, recall that what Bostrom (2014, p.93) terms "superpowers" are contextual, being properties not of agents *per se*,[4] but of *agents that have an effective monopoly on the capability in question* (Bostrom 2014, p.104): In a prepared world, mere superintelligence would not confer superpowers. (Regarding the "hacking superpower" (Bostrom 2014, p.94), note that, even today, practical operating systems can provide mathematically provable, hence unhackable, security guarantees.[5])

32.5 Strong defensive capabilities can constrain problematic agents

The above points offer only an abstract sketch of a development process and objective, not a map of a road or a destination. A closer look can help to clarify key concepts:

---

1. Section 26: Superintelligent-level systems can safely provide design and planning services
2. Section 27: Competitive pressures provide little incentive to transfer strategic decisions to AI systems
3. Section 23: AI development systems can support effective human guidance
4. Section 31: Supercapabilities do not entail "superpowers"
5. `https://sel4.systems/`

1) Preparatory, SI-level red-team/blue-team design competition can explore potential attacks while exploring the conditions necessary for security services to block attack capabilities with an ample margin of safety. Adversarial exercises could readily employ physical simulations that are *qualitatively* biased to favor hypothetical attackers, while assigning arbitrarily large, highly-asymmetric *quantitative* advantages to proposed security services. As noted above, enormous resources could potentially be mobilized to support SI-level exploration of hypothetical red-team threats and proposed blue-team security measures; thorough exploration would call for a good working approximation to what Bostrom (2014, p.229) terms "technological completion", at least in a design sense.

2) Anticipatory deployment of well-resourced, SI-level security services would implement systems that reflect the results of stringent red-team/blue-competitions, and hence would employ more-than-adequate physical and computational resources. Note that preparatory, selective development and deployment of security systems strongly embodies what Bostrom (2014, p.230) terms "differential technology development".

3) Ongoing application of effective, physically-oriented surveillance calls for collection of information sufficient to establish reliable (yet not excessively conservative) upper bounds on the scope of potentially threatening physical capabilities of potentially untrustworthy actors. Recognition of threats can be informed by risk-averse generalizations of worst-case red-team strategies.

4) Ongoing application of effective, physically-oriented security measures calls for the application of ample (yet not unnecessarily conservative) resources to forestall potential threats; policies can be informed by amply (yet not excessively conservative) risk-averse generalizations of robust blue-team security measures. Crude security measures might require either strong interventions or stringent constraints on actors' physical resources; well-designed security measures could presumably employ milder interventions and constraints, optimized for situation-dependent acceptability conditioned on global effectiveness.

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

The concept of impartial and effective AI-enabled security services raises questions regarding the deep underpinnings of a desirable civilizational order, questions that cannot be explored without raising raising further questions at levels that range from security policies and physical enforcement to the entrenchment of constitutional orders and the potential diversity of coexisting frameworks of law. Prospects for a transition to a secure, AI-stable world raise further questions regarding potential paths forward, questions that involve not only technological developments, but ways in which the perceived interests and options of powerful, risk-averse actors might align well enough to shape actions that lead to widely-approved outcomes.

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

Discussions of superintelligence and AI safety often envision the development of an extremely powerful AI agent that will take control of the world and optimize the future in accord with human values. This scenario presents several difficulties: It seems impossible to define human values in a way that would be generally accepted, impossible to implement systems that would be trusted to optimize the world, and difficult to take control of the world (whether openly or by attempted stealth) without provoking effective, preemptive opposition from powerful actors. Fortunately, as outlined above, the foundational safety challenge—physical security—can be addressed while avoiding these problems.

Further Reading

- *Section 12: AGI agents offer no compelling value*
- *Section 20: Collusion among superintelligent oracles can readily be avoided*
- *Section 23: AI development systems can support effective human guidance*
- *Section 26: Superintelligent-level systems can safely provide design and planning services*
- *Section 27: Competitive pressures provide little incentive to transfer strategic decisions to AI systems*
- *Section 31: Supercapabilities do not entail "superpowers"*

33 Competitive AI capabilities will not be boxed

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

Current trends suggest that superintelligent-level AI capabilities will emerge from a distributed, increasingly automated process of AI research and development, and it is difficult to envision a scenario in which a predominant portion of overall AI capacity would (or could) emerge and be confined in "a box". Individual systems could be highly problematic, but we should expect that AI systems will exist in a milieu that enables instantiation of diverse peer-level systems, a capability that affords scalable, potentially effective mechanisms for managing threatening AI capabilities.

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

Current trends in machine learning point to an increasing range of superhuman capabilities emerging from extensions of today's technology base that emerge from extensions of today's R&D milieu. Today we see a distributed, increasingly automated R&D process that employs a heterogeneous toolset to develop diverse demonstration, prototype, and application systems. Increasingly, we find that AI applications include AI development tools, pointing the way toward thorough automation of development that would enable AI progress at AI speed—an incremental model of asymptotically recursive technology improvement.[1]

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

Current methods in machine learning suggest that access to large-scale machine resources will be critical to competitive performance in AI technology and applications development. The growing diversity of technologies and applications (*e.g.*, to vision, speech recognition, language translation, ML

1. Section 10: R&D automation dissociates recursive improvement from AI agency

architecture design. . .) speak against the idea that a world-class R&D process (or its functional equivalent) will someday be embodied a distinct, general-purpose system. In other words, it is difficult to envision plausible scenarios in which we find more AI capability "in a box" than in the world outside.

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

We should expect that any particular AI system will be embedded in an extended AI R&D ecosystem having aggregate capabilities that exceed its own. Any particular AI architecture will be a piece of software that can be trained and run an indefinite number of times, providing multiple instantiations that serve a wide range of purposes (a very wide range of purposes, if we posit truly general learning algorithms). As is true today, we can expect that the basic algorithms and implementation techniques that constitute any particular architecture will be deployed in diverse configurations, trained on diverse data, and provide diverse services.[1]

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

Absent systemic constraints, advanced AI technologies will enable the implementation of systems that are radically unsafe or serve abhorrent purposes. These prospects can be classed as bad-actor risks that, in this framing, include actions that incur classic AGI-agent risks.

The almost unavoidable ability to instantiate diverse AI systems at any given level of technology also offers benefits for AI reliability and safety. In particular, the ability to instantiate diverse peer-level AI systems enables the use of architectures that rely on implicitly competitive and adversarial relationships among AI components, an approach that enables the use of AI systems to manage other AI systems while avoiding concerns regarding potential collusion.[2] Both competitive and adversarial mechanisms are found in current AI practice, and scale to a superintelligent level.

Further Reading

- *Section 10: R&D automation dissociates recursive improvement from AI agency*

1. Section 15: Development-oriented models align with deeply-structured AI systems
2. Section 20: Collusion among superintelligent oracles can readily be avoided

- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 20: Collusion among superintelligent oracles can readily be avoided*

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

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

AI R&D is currently distributed across many independent research groups, and the architecture of R&D automation is compatible with continued decentralized development. Various pressures tend to favor greater centralization of development in leading organizations, yet centralization *per se* would neither force nor strongly favor a qualitative change in the architecture of R&D tasks. Alternative distributions of capabilities across organizations could provide affordances relevant to AI policy and strategy.

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

State-of-the-art AI research and development is currently decentralized, distributed across independent groups that operate within a range of primarily corporate and academic institutions. Continued automation of AI R&D tasks[1] will likely increase the advantages provided by proprietary tool-sets, integrated systems, and large-scale corporate resources, yet strong automation is compatible with continued decentralization.

34.3 Accelerating progress could lead to strong centralization of capabilities

Fundamental arguments suggest that AI R&D automation provides the most direct path[2] to steeply accelerating, AI-enabled progress in AI technologies.

1. Section 10: R&D automation dissociates recursive improvement from AI agency
2. Section 1: R&D automation provides the most direct path to an intelligence explosion

Steeply accelerating progress, if driven by proprietary, rapidly-advancing tool sets, could favor the emergence of wide gaps between competing groups, effectively centralizing strong capabilities in a leader.

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

Pressures that favor centralization neither force nor strongly favor a qualitative change in the architecture of R&D tasks or their automation. Organizational centralization, tool-chain integration, and task architecture are distinct considerations, and only loosely coupled.

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

Considerations involving AI policy and strategy may favor centralization of strong capabilities (*e.g.*, to provide affordances for centralized control), or might favor the division of complementary capabilities across organizations (*e.g.*, to provide affordances for establishing cross-institutional transparency and interdependence). Unlike classic models of advanced AI capabilities as something embodied in a distinct entity ("the machine"), the R&D automation model is compatible with both alternatives.

Further Reading

- *Section 1: R&D automation provides the most direct path to an intelligence explosion*

35 Predictable aspects of future knowledge can inform AI safety strategies

> 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

Along realistic development paths, researchers building advanced AI systems will gain extensive safety-relevant knowledge from experience with similar but less advanced systems. While we cannot predict the specific content

of that future knowledge, we can have substantial confidence regarding its scope; for example, researchers will have encountered patterns of success and failure in both development and applications, and will have eagerly explored and exploited surprising capabilities and behaviors across multiple generations of AI technology. Realistic models of potential AI development paths and risks should anticipate that this kind of knowledge will be available to contemporaneous decision makers, hence the nature and implications of future safety-relevant knowledge call for further exploration by the AI safety community.

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

Although we cannot predict the details of future AI technologies and systems, we can predict that their developers will know more about those systems than we do. In general, the nature of knowledge learned during technology development is strictly more predictable than the content of the knowledge itself, hence we can consider the expected scope of future known-knowns and known-unknowns, and even the expected scope of knowledge regarding unknown-unknowns—expected knowledge of patterns of ongoing surprises. Thus, in studying AI safety, it is natural to consider not only our current, sharply limited knowledge of future technologies, but also our somewhat more robust knowledge of the *expected scope* of future knowledge, and of the expected scope of *future knowledge regarding expected surprises.* These are aspects of *anticipated contemporaneous knowledge.*

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

By design (and practical necessity), low-power nuclear chain reactions preceded nuclear explosions, and despite best efforts, small aircraft crashed before large aircraft. In AI, successes and failures of MNIST classification preceded successes and failures of ImageNet classification, which preceded successes and failures of machine vision systems in self-driving cars.

In particular, we can expect that future classes of AI technologies that could yield *enormously* surprising capabilities will already have produced *impressively* surprising capabilities; with that experience, to encounter outliers—surprising capabilities of enormous magnitude—would not be enormously surprising.

35.4 AI researchers eagerly explore and exploit surprising capabilities

Scientists and researchers focused on basic technologies are alert to anomalies and strive to characterize and understand them. Novel capabilities are pursued with vigor, and unexpected capabilities are celebrated: Notable examples in recent years include word embeddings that enable the solution of word-analogy problems by vector arithmetic, and RL systems that learn the back-cavity multiple-bounce trick in Atari's Breakout game. Surprising capabilities will be sought, and when discovered, they will be tested and explored.

35.5 AI developers will be alert to patterns of unexpected failure

Technology developers play close attention to performance: They instrument, test, and compare alternative implementations, and track patterns of success and failure. AI developers seek low error rates and consistent performance in applications; peculiarities that generate unexpected adverse results are (and will be) studied, avoided, or tolerated, but not ignored.

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

Several years ago, one could imagine that AI safety concerns might be ignored by AI developers, and it was appropriate to ask how safety-oblivious development might go awry. AI safety concerns, however, led to the growth of AI safety studies, which are presently flourishing. We can expect that safety studies will be active, ongoing, and substantially integrated with the AI R&D community, and will be able to exploit contemporaneous community knowledge in jointly developing and revising practical recommendations.

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

Adoption of safety-oriented recommendations will depend in part on their realism and practicality, considerations that call for a better understanding of the potential extent of future safety-relevant in the development community. Studies of the nature of past technological surprises can inform this effort, as can studies of patterns of development, anticipation, and surprise in modern AI research.

We must also consider potential contrasts between past patterns of development and future developments in advanced AI. If contemporaneous computational hardware capacity will (given the development of suitable software) be sufficient to support broadly superhuman performance,[1] then the potential for swift change will be unprecedented. Contingent on informed caution and a security mindset, however, the *potential* for swift change need not entail unsafe application of capabilities and unprecedented, unavoidable surprises. To understand how such a situation might be managed, it will be important to anticipate the growth of safety-relevant knowledge within the AI development community, and to explore how this knowledge can inform the development of safety-oriented practices.

Further Reading

- *Section 12: AGI agents offer no compelling value*
- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 16: Aggregated experience and centralized learning support AI-agent applications*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*

36 Desiderata and directions for interim AI safety guidelines

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

Actionable, effective interim AI safety guidelines should:

- Clarify why current AI research is safe,
- Promote continued safety-enabling development practice, and
- Foster ongoing, collaborative guideline development and adoption.

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

Because current AI development work is, in fact, safe with respect to high-level risks, interim safety guidelines can clarify and codify safety-aligned characteristics of current work while placing little burden on practitioners. Good practice in current AI R&D tends to align with longer-term safety concerns: Examples include learning from the exploration of families of architectures and tasks, then pursuing task-oriented development, testing, and validation before building complex deployed systems. These practices can contribute to shaping and controlling AI capabilities across a range of potential development paths. Development and adoption of guidelines founded on current practice could help researchers answer public questions about the safety of their work while fostering ongoing, collaborative safety research and guideline extension to address potential longer-term, high-level risks.

36.2 Desiderata

36.2.1 *Desideratum:* Clarify why current AI research is safe

Current AI research is safe (in a classic x-risk sense), in part because current AI capabilities are limited, but also because of the way capabilities are developed and organized. Interim guidelines could clarify and codify aspects of current practice that promote foundational aspects of safety (see below), and thereby support efforts to identify safe paths to more powerful capabilities.

36.2.2 *Desideratum:* Promote continued safety-enabling development practice

Guidelines that focus on safety-promoting aspects of current practice can be crafted to place little burden on what is already safety-compliant research; these same safety-promoting practices can contribute to (though not in themselves ensure) avoiding hazards in more challenging future situations.

36.2.3 *Desideratum:* Foster ongoing, collaborative guideline development and adoption

Collaboration on actionable interim safety guidelines could promote closer links between development- and safety-oriented AI researchers, fostering ongoing collaborative, forward-looking guideline development. Beginning with readily-actionable guidelines can help to ensure that collaboration goes beyond theory and talk.

36.3 Good practice in development tends to align with safety concerns

Fortunately (though unsurprisingly) good practice in AI development tends to align with safety concerns. In particular, developers seek to ensure that AI systems behave predictably, a characteristic that contributes to safety even when imperfect.

36.4 Exploring families of architectures and tasks builds practical knowledge

In AI R&D, we see extensive exploration of families of architectures and tasks through which developers gain practical knowledge regarding the capabilities and (conversely) limitations of various kinds of systems; practical experience also yields an understanding of the kinds of surprises to be expected from these systems. Guidelines that highlight the role of present and future practical knowledge[1] would clarify why current research is known to be safe, and how good development practice can contribute to future safety.

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

Systems for practical applications perform bounded tasks and are subject to testing and validation before deployment. Task-oriented development, testing, and validation contribute to knowledge of capabilities and focus strongly-motivated attention on understanding potential failures and surprises. Guidelines that codify this aspect of current practice would again help to clarify conditions that contribute to current and future safety.

36.6 Modular architectures make systems more understandable and predictable

Systems are more easily designed, developed, tested, and upgraded when they are composed of distinct parts, which is to say, when their architectures are modular rather than opaque and undifferentiated. This kind of structure is ubiquitous in complex systems. (And as noted in a recent paper from Google, "Only a small fraction of real-world ML systems is composed of the ML code [...] The required surrounding infrastructure is vast and complex." [Sculley

1. Section 35: Predictable aspects of future knowledge can inform AI safety strategies

et al. 2015]) In particular, the distinction between AI development systems and their products[1] enables a range of reliability (hence safety) oriented practices.[2] The use of modular architectures in current practice again suggests opportunities for codification, explanation, and contributions to future safety.

36.7 Interim safety guidelines can foster ongoing progress

The absence of current safety risks sets a low bar for the effectiveness of interim guidelines, yet guidelines organized around current practice can contribute to the development of guidelines that address more challenging future concerns. At a technical level, practices that support reliability in narrow AI components can provide foundations for the safe implementation of more capable systems. At an institutional level, linking current practice to longer-term concerns can foster safety-oriented research and development in several ways: by encouraging understanding and extension of what today constitutes good practice, by engaging the development community in ongoing guideline development, and by focusing greater research attention on the potential connections between development processes and safe outcomes. Interim guidelines cannot solve all problems, yet could help to set our work on a productive path.

Further Reading

- *Section 10: R&D automation dissociates recursive improvement from AI agency*
- *Section 12: AGI agents offer no compelling value*
- *Section 22: Machine learning can develop predictive models of human approval*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 35: Predictable aspects of future knowledge can inform AI safety strategies*

1. Section 15: Development-oriented models align with deeply-structured AI systems
2. Section 16: Aggregated experience and centralized learning support AI-agent applications

37 How do neural and symbolic technologies mesh?

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

Neural network (NN) and symbolic/algorithmic (S/A) AI technologies offer complementary strengths, and these strengths can be combined in multiple ways. A loosely-structured taxonomy distinguishes several levels of organization (components, algorithms, and architectures), and within each of these, diverse modes of integration—various functional relationships among components, patterns of use in applications, and roles in architectures. The complexity and fuzziness of the taxonomy outlined below emphasizes the breadth, depth, and extensibility of current and potential NN–S/A integration.

37.2 Motivation

One might imagine that neural network and symbolic/algorithmic technologies are in competition, and ask whether NNs can fulfill the grand promise of artificial intelligence when S/A methods have failed—will NN technologies also fall short?

On closer examination, however, the situation looks quite different: NN and S/A technologies are not merely in competition, they are complementary, compatible, and increasingly integrated in research and applications. To formulate a realistic view of AI prospects requires a general sense of the relationship between NN and S/A technologies. Discussions in this area are typically more narrow: They either focus on a problem domain and explore applicable NN–S/A techniques, or they focus on a technique and explore potential applications. The discussion here will instead outline the *expanding range of techniques and applications*, surveying patterns of development that may help us to better anticipate technological opportunities and the trajectory of AI development.

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

There is no sharp and natural criterion that distinguishes NN from S/A techniques. For purposes of discussion, one can regard a technique as NN-style to the extent that it processes numerical, semantically-opaque vector representations through a series of transformations in which operations and data-flow patterns are fixed. Conversely, one can regard a technique as S/A-style to the extent that it relies on entities and operations that have distinct meanings and functions, organized in space and time in patterns that manifestly correspond to the structure of the problem at hand—patterns comprising data structures, memory accesses, control flow, and so on.

Note that S/A *implementation mechanisms* should not be mistaken for S/A *systems:* S/A code can implement NN systems much as hardware implements software, and just as code cannot usefully be reduced to hardware, so NNs cannot usefully be reduced to code. In the present context, however, the systems of greatest interest are those that deeply integrate NN- and S/A-style mechanisms, or that blur the NN–S/A distinction itself. If taxonomy were clean and easily constructed, prospects for NN–S/A integration would be less interesting.

37.4 NN and S/A techniques are complementary

The contrasting strengths of classic S/A techniques and emerging NN techniques are well known: S/A-style AI techniques have encountered difficulties in perception and learning, areas in which NN techniques excel; NN–style AI techniques, by contrast, often struggle with tasks like logic-based reasoning—and even counting—that are trivial for S/A systems.

In machine translation, for example, algorithms based on symbolic representations of syntax and semantics fell short of their promise (even when augmented by statistical methods applied to large corpora); more recently, neural machine translation systems with soft, opaque representations have taken the lead, yet often struggle with syntactic structure and the semantics of logical entailment.

S/A systems are relatively transparent, in part because S/A systems embody documentable, human-generated representations and algorithms, while NN systems instead discover and process opaque representations in ways that do not correspond to interpretable algorithms. The integration of NN techniques with S/A systems can sometimes facilitate interpretability, however: For example, NN operations may be more comprehensible when considered

as functional blocks in S/A architectures, while (in sometimes amounts to a figure-ground reversal) comprehensible S/A systems can operate on NN outputs distilled into symbols or meaningful numerical values.

37.5 AI-service development can scale to comprehensive, SI-level services

In discussing how NN and S/A techniques mesh, it will be convenient to draw rough distinctions between three levels of application:

Components and mechanisms, where NN and S/A building blocks interact at the level of relatively basic programming constructs (*e.g.*, data access, function calls, branch selection).

Algorithmic and representational structures, where NN and S/A techniques are systematically intwined to implement complex representations or behaviors (*e.g.*, search, conditional computation, message-passing algorithms, graphical models, logical inference).

Systems and subsystems, where individually-complex NN and S/A subsystems play distinct and complementary roles at an architectural level (*e.g.*, perception and reasoning, simulation and planning).

The discussion below outlines several modes of NN–S/A integration at each level, each illustrated by an unsystematic sampling of examples from the literature. No claim is made that the modes are sharply defined, mutually exclusive, or collectively exhaustive, or that particular examples currently outperform alternative methods. The focus here is on patterns of integration and proofs of concept.

37.6 Integration at the level of components and mechanisms

What are conceptually low-level components may have downward- or upward-facing connections to complex systems: Components that perform simple functions may encapsulate complex NN or S/A mechanisms, while simple functions may serve as building blocks in higher-level algorithmic and representational structures (as discussed in the following section).

37.6.1 NNs can provide representations processed by S/A systems:

Classic AI algorithms manipulate representations based on scalars or symbolic tokens; in some instances (discussed below), systems can retain the architecture of these algorithms—patterns of control and data flow—while exploiting richer NN representations. For example, the modular, fully-differentiable visual question answering architecture of (Hu et al. 2018) employs S/A-style mechanisms (sets of distinct, compositional operators that pass data on a stack), but the data-objects are patterns of soft attention over an image.

37.6.2 NNs can direct S/A control and data flow:

In AI applications, S/A algorithms often must select execution paths in an "intelligent" way. NNs can process rich information (*e.g.*, large sets of conventional variables, or NN vector embeddings computed upstream), producing Boolean or integer values that can direct these S/A choices. NN-directed control flow is fundamental to the NN-based search and planning algorithms noted below.

37.6.3 S/A mechanisms can direct NN control and data flow:

Conversely, S/A mechanisms can structure NN computations by choosing among alternative NN components and operation sequences. Conditional S/A-directed NN operations enable a range of NN–S/A integration patterns discussed below.

37.6.4 NNs can learn heuristics for S/A variables:

Reinforcement learning ML mechanisms be applied to what are essentially heuristic computations (*e.g.*, binary search, Quicksort, and cache replacement) by computing a value base on a observations (the values of other variables). This approach embeds ML in a few lines of code to "integrate ML tightly into algorithms whereas traditional ML systems are build around the model" (Carbune et al. 2017).

37.6.5 NNs can replace complex S/A functions:

Conventional algorithms may call functions that perform costly numerical calculations that give precise results when approximate results would suffice. In the so-called "parrot transformation" (Esmaeilzadeh et al. 2012), an NN is trained to mimic and replace the costly function. NNs that (approximately)

model the trajectories of physical systems (Ehrhardt et al. 2017) could play a similar role by replacing costly simulators.

37.6.6 NNs can employ complex S/A functions:

Standard deep learning algorithms learn by gradient descent on linear vector transformations composed with simple, manifestly-differentiable element-wise functions (ReLU, tanh, *etc.*), yet complex, internally non-differentiable algorithms can also implement differentiable functions. These functions can provide novel functionality: For example, a recent deep-learning architecture (OptNet) treats constrained, exact quadratic optimization as a layer, and can learn to solve Sudoku puzzles from examples (Amos and Kolter 2017). When considered as functions, the outputs of complex numerical models of physical systems can have a similar differentiable character.

37.6.7 S/A algorithms can extend NN memory:

Classic NN algorithms have limited representational capacity, a problem that becomes acute for recurrent networks that must process long sequences of inputs or represent an indefinitely large body of information. NNs can be augmented with scalable, non-differentiable (hard-attention) or structured memory mechanisms (Sukhbaatar et al. 2015; Chandar et al. 2016) that enable storage and retrieval operations in an essentially S/A style.

37.6.8 S/A data structures can enable scaling of NN representations:

Data structures developed to extend the scope of practical representations in S/A systems can be adapted to NN systems. Hash tables and tree structures can support sparse storage for memory networks, for example, and complex data structures (octrees) enable generative convolutional networks to output fine-grained 3D representations that would otherwise require impractically large arrays (Tatarchenko, Dosovitskiy, and Brox 2017); access to external address spaces has proved critical to solving complex, structured problems (Graves et al. 2016). Brute-force nearest-neighbor lookup (*e.g.*, in NN embedding spaces) is widely used in single-shot and few-shot learning (see below); recent algorithmic advances based on neighbor graphs enable retrieval of near neighbors from billion-scale data sets in milliseconds (Fu, Wang, and Cai 2017).

37.6.9 S/A mechanisms can template NN mechanisms:

In a more abstract relationship between domains, S/A-style mechanisms can be implemented in a wholly differentiable, NN form. Examples include pointer networks (Vinyals, Fortunato, and Jaitly 2015) that (imperfectly) solve classic S/A problems such as Traveling Salesman and Delaunay triangulation, as well as differentiable stacks that can perform well on NLP problems (e.g., syntactic transformations [Grefenstette et al. 2015] and dependency parsing [Dyer et al. 2015]) commonly addressed by recursive, tree-structured algorithms in S/A systems.

37.7 Integration at the level of algorithmic and representational structures

Higher-level integration of NN and S/A mechanisms can typically be viewed as patterns of interleaved NN and S/A operations, sometimes with substantial, exposed complexity in one or both components.

37.7.1 S/A algorithms can extend NN inference mechanisms:

S/A mechanisms are often applied to decode NN outputs. For example, beam search in is a classic algorithm in symbolic AI, and is applied in neural machine translation systems to select sequences of symbols (*e.g.*, words) based on soft distributions over potential sequence elements. In another class of algorithms, S/A algorithms are applied to find near-neighbors among sets of vector outputs stored in external memory, supporting one- and few-shot learning in classifiers (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017).

37.7.2 NNs can guide S/A search:

Both search over large spaces of choices (Go play) and large bodies of data (the internet) and are now addressed by traditional S/A methods (Monte Carlo tree search [Silver et al. 2016], large-scale database search [Clark 2015]) guided by NN choices. Recent work has integrated tree-based planning methods with NN "intuition" and end-to-end learning to guide agent behaviors (Anthony, Tian, and Barber 2017; Farquhar et al. 2017; Guez et al. 2018).

37.7.3 S/A graphical models can employ NN functionality:

An emerging class of graphical models—message-passing NNs—represents both node states and messages as vector embeddings. Message-passing NNs

share the discrete, problem-oriented structures of probabilistic graphical models, but have found a far wider range of applications, including the prediction of molecular properties (Gilmer et al. 2017), few-shot learning (Garcia and Bruna 2017), and inferring structured representations (Johnson et al. 2016), as well as algorithms that outperform conventional loopy belief propagation (Yoon et al. 2018), and others that can *infer causality from statistical data* beyond the limits that might be suggested by Pearl's formalism (Goudet et al. 2017). Vicarious has demonstrated integration of perceptual evidence through a combination of NN-level pattern recognition and loopy graphical models: "Recursive Cortical Networks" can generalize far better than conventional NNs, and from far less training data (George et al. 2017).

37.7.4 S/A mechanisms can structure NN computation:

S/A mechanisms are used both to construct specified graphs and to execute corresponding message-passing algorithms, while *specification* of graph-structured NNs is a task well-suited to mixed NN and S/A mechanisms. For example, S/A and NN mechanisms have been interleaved to compose and search over alternative tree-structured NNs that implement generative models of 3D structures (Jun Li et al. 2017). In a very different example (a visual question-answering task), an S/A parsing mechanism directs the assembly of question-specific deep networks from smaller NN modules (Hu et al. 2018); see also Andreas et al. (2016). A similar strategy is employed in the Neural Rule Engine (Li, Xu, and Lu 2018), while Yi et al. (2018) employs NN mechanisms to parse scenes and questions into symbolic representations that drive a symbolic execution engine.

37.7.5 NNs can produce and apply S/A representations:

NN mechanisms can learn discrete, symbol-like representations useful in reasoning, planning, and predictive learning (van den Oord, Vinyals, and Kavukcuoglu 2017; Raiman and Raiman 2018), including string encodings of molecular-structure graphs for chemical synthesis (Segler, Preuß, and Waller 2017) and drug discovery (Merk et al. 2018). S/A representations can enforce a strong inductive bias toward generalization; for example, deep RL systems can formulate and apply symbolic rules (Garnelo, Arulkumaran, and Shanahan 2016), and NN techniques can be combined with inductive logic programming to enable the inference of universal rules from noisy data (Evans and Grefenstette 2018).

37.7.6 S/A algorithms can template NN algorithms:

S/A representations can readily implement geometric models with parts, wholes, and relationships among distinct objects. In recent NN architectures, "capsules" (Sabour, Frosst, and Hinton 2017) and similar components (Liao and Poggio 2017) can both support image recognition and play a symbol-like role in representing part-whole relationships; architectures that embody relation-oriented priors can both recognize objects and reason about their relationships (Hu et al. 2018) or model their physical interactions (Battaglia et al. 2016; Chang et al. 2016). The neural architectures that accomplish these tasks follow (though not closely!) patterns found in S/A information processing. More broadly, there is a recognized trend of learning differentiable versions of familiar algorithms (Guez et al. 2018).

37.7.7 NNs can learn S/A algorithms:

In neural program induction, NNs are trained to replicate the input-output behavior of S/A programs. Several architectures (e.g., the Neural Turing Machine [Graves, Wayne, and Danihelka 2014], Neural GPU [Kaiser and Sutskever 2015], and Neural Programmer [Neelakantan, Le, and Sutskever 2015]; reviewed in Kant [2018]) have yielded substantial success on simple algorithms, with (usually imperfect) generalization to problem-instances larger than those in the training set.

37.7.8 NNs can aid S/A program synthesis:

Automatic programming is a long-standing goal in AI research, but progress has been slow. Mixed S/A–NN methods have been applied to program synthesis (Yin and Neubig 2017; Singh and Kohli 2017; Abolafia et al. 2018), with increasing success (reviewed in (Kant 2018)). Potential low-hanging fruit includes adaptation of existing source code (Allamanis and Brockschmidt 2017) and aiding human programmers by code completion (Jian Li et al. 2017); automating the development of glue code would facilitate integration of existing S/A functionality with other S/A and NN components.

We can anticipate that advances in graph NNs will facilitate the exploitation of richer structures in code, which has traditionally centered on narrower syntactic representations. Structures latent in S/A code, but to date not explored *in conjunction*, include not only abstract syntax trees, data types, and function types, but control and data flow graphs *(May 2018 update: see Allamanis, Brockschmidt, and Khademi [2018]).*

37.7.9 NNs can aid automated theorem proving:

Automated theorem proving systems perform heuristic search over potential proof trees, and deep-learning methods have been applied to improve premise selection (Irving et al. 2016; Kaliszyk, Chollet, and Szegedy 2017). Progress in automated theorem proving (and proof assistants) could facilitate the development of provably correct programs and operating systems (Klein et al. 2014).

37.7.10 S/A mechanisms can support NN architecture and hyperparameter search:

New tools for architecture and hyperparameter search are accelerating NN development by discovering new architectures (Zoph and Le 2016; Pham et al. 2018) and optimizing hyperparameters (Jaderberg et al. 2017) (a key task in architecture development). Leading methods in architecture search apply NN or evolutionary algorithms to propose candidate architectures, while using an S/A infrastructure to construct and test them.

37.8 Integration at the level of systems and subsystems

Integration at the level of systems and subsystems extends the patterns already discussed, combining larger blocks of NN and S/A functionality.

37.8.1 NNs can support and ground S/A models:

Perceptual processing has been chronically weak in S/A artificial intelligence, and NN techniques provide a natural complement. NN-based machine vision applications in robotics and vehicle automation are expanding (Steger, Ulrich, and Wiedemann 2007), and NN-based modeling can go beyond object recognition by, for example, enabling the inference of physical properties of objects from video (Watters et al. 2017). With the integration of NN perception and symbolic representations, symbol-systems can be grounded.

37.8.2 S/A representations can direct NN agents:

In recent work, deep RL has been combined with symbolic programs, enabling the implementation of agents that learn correspondences between programs, properties, and objects through observation and action; these capabilities can be exploited in systems that learn to ground and execute explicit, human-written programs (Denil et al. 2017). The underlying principles should gen-

eralize widely, improving human abilities to inform, direct, and understand behaviors that exploit the strengths of deep RL.

37.8.3 NNs can exploit S/A models and tools:

Looking forward, we can anticipate that NN systems, like human beings, will be able to employ state-of-the-art S/A computational tools, for example, using conventional code that implements physical models, image rendering, symbolic mathematics and so on. NN systems can interact with S/A systems though interfaces that are, at worst, like those we use today.

37.8.4 NN and S/A models can be integrated in cognitive architectures:

At a grand architectural level, AI researchers have long envisioned and proposed "cognitive architectures" (Soar, LIDA, ACT-R, CLARION...) intended to model much of the functionality of the human mind. A recent review (Kotseruba, Gonzalez, and Tsotsos 2016) identifies 84 such architectures, including 49 that are still under active development. Recent work in cognitive architectures has explored the integration of S/A and NN mechanisms (Besold et al. 2017), an approach that could potentially overcome difficulties that have frustrated previous efforts.

37.9 Integration of NN and S/A techniques is a rich and active research frontier

An informal assessment suggests robust growth in the literature on integration of NN and S/A techniques. It is, however, worth noting that there are incentives to focus research efforts primarily on one or the other. The most obvious incentive is intellectual investment: Crossover research requires the application of disparate knowledge, while more specialized knowledge is typically in greater supply for reasons of history, institutional structure, and personal investment costs. These considerations tend to suggest an undersupply of crossover research.

There is, however, good reason to focus extensive effort on NN systems that that *do not* integrate S/A techniques in a strong, algorithmic sense: We do not yet know the limits of NN techniques, and research that applies NNs in a relatively pure form—end-to-end, *tabula rasa* training with S/A code providing only infrastructure or framework elements—seems the best way to explore the NN frontier. Even if one expects integrated systems to dominate the world

of applications, relatively pure NN research may be the most efficient way to develop the NN building blocks for those applications.

As in many fields of endeavour, it is important to recognize the contrasts between effective methodologies in research and engineering: In particular, good basic research explores systems that are novel, unpredictable, and (preferably) simple, while good engineering favors known, reliable building blocks to construct systems that are as complex as a task may require. Accordingly, as technologies progress from research to applications, we can expect to see increasing—and increasingly eclectic—integration of NN and S/A techniques, providing capabilities that might otherwise be beyond our reach.

Further Reading

- *Section I: Introduction: From R&D automation to comprehensive AI Services*
- *Section II: Overview: Questions, propositions, and topics*

38 Broadly-capable systems coordinate narrower systems

> 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

In today's world, superhuman competencies reside in structured organizations with extensive division of knowledge and labor. The reasons for this differentiated structure are fundamental: Specialization has robust advantages both in learning diverse competencies and in performing complex tasks. Unsurprisingly, current AI services show strong task differentiation, but perhaps more surprisingly, AI systems trained on seemingly indivisible tasks (*e.g.*, translating sentences) can spontaneously divide labor among "expert" components. In considering potential SI-level AI systems, black-box abstractions may sometimes be useful, but these abstractions set aside our general knowledge of the differentiated architecture of intelligent systems.

38.2 Today's superhuman competencies reside in organizational structures

It is a truism that human organizations can achieve tasks beyond individual human competence by employing, not just many individuals, but individuals who perform differentiated tasks using differentiated knowledge and skills. Adam Smith noted the advantages of division of labor (even in making pins), and in modern corporations, division of labor among specialists is a necessity.

38.3 Specialization has robust advantages in learning diverse competencies

The structure of knowledge enables parallel training: Learning competencies in organic chemistry, financial management, mechanical engineering, and customer relations, for example, is accomplished by individuals who work in parallel to learn the component tasks. This pattern of parallel, differentiated learning works well because many blocks of specialized knowledge have little mutual dependence. The limited pace of human learning and vast scope of human knowledge make parallel training mandatory in the human world, and in machine learning analogous considerations apply. Loosely-coupled bodies of knowledge call for loosely-coupled learning processes that operate in parallel.

38.4 Division of knowledge and labor is universal in performing complex tasks

As with learning, parallel, specialized efforts have great advantages in performing tasks. Even setting aside human constraints on bandwidth and representational power, there would be little benefit in attempting to merge day-to-day tasks in the domains of organic chemistry, financial management, mechanical engineering, and customer relations. Both information flow and knowledge naturally cluster in real-world task structures, and the task of cross-task management (*e.g.*, in developing and operating a chemical-processing system) has only limited overlap with the information flows and knowledge that are central to the tasks that must be coordinated. To implement a complex, inherently differentiated task in a black-box system is to reproduce the task structure inside the box while making its organization opaque.

38.5 Current AI services show strong task differentiation

It goes without saying that current AI systems are specialized: Generality is a challenge, not a default. Even in systems that provide strong generalization *capacity*, we should expect to see diminishing returns from attempts to apply *single-system* generalization capacity to the full scope and qualitative diversity of human knowledge. Indeed, considering the nature of training and computation, it is difficult to imagine what "single-system generalization" could even mean on that scale.

38.6 AI systems trained on seemingly indivisible tasks learn to divide labor

Task specialization can emerge spontaneously in machine learning systems. A striking recent example is a mixture-of-experts model employed in a then state-of-the art neural machine translation system to enable 1000x improvements in model capacity (Shazeer et al. 2017). In this system, a managerial component delegates the processing of sentence fragments to "experts" (small, architecturally undifferentiated networks), selecting several from a pool of thousands. During training, experts spontaneously specialize in peculiar, semantically- and syntactically-differentiated aspects of text comprehension. The incentives for analogous specialization and task delegation can only grow as tasks become wider in scope and less tightly coupled.

38.7 Black-box abstractions discard what we know about the architecture of systems with broad capabilities

Appropriate levels of abstraction depend on both our knowledge and our purpose. If we want to model the role of Earth in the dynamics of the Solar System, it can be treated as a point mass. If we want to model Earth as a context for humanity, however, we also care about its radius, geography, geology, climate, and more—and we have substantial, useful knowledge of these. Likewise, for some purposes, it can be appropriate to model prospective AI systems as undifferentiated, black-box pools of capabilities. If we want to understand prospective AI systems in the context of human society, however, we have strong practical reasons to apply what we know about the general architecture of systems that perform broad tasks.

Further Reading

- *Section 12: AGI agents offer no compelling value*

- *Section 15: Development-oriented models align with deeply-structured AI systems*
- *Section 21: Broad world knowledge can support safe task performance*
- *Section 28: Automating biomedical R&D does not require defining human welfare*
- *Section 39: Tiling task-space with AI services can provide general AI capabilities*

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

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

Task-centered models of general AI capabilities highlight the importance of matching tasks to services, and current practice in deep learning suggests both conceptual models and concrete approaches. A surprisingly wide range of operations in deep-learning systems link "tasks" to "services" through what are—or can be construed as—proximity-based access operations in high-dimensional vector embedding spaces. Applications of proximity-based access include single-shot learning, situational memory in RL agents, mixture-of-experts models, and matching human queries to physical products across billion-scale product databases. In light of these diverse applications, it is natural to consider proximity in embedding spaces as a basis for scalable access to functional components at levels ranging from fine-grained perceptual tasks to integrated, high-level services visible to human users. Similar mechanisms could facilitate the implementation of new services both through adaptation of existing services and through development of new systems for novel tasks. Services can be seen tiles covering regions of task-space: Services of greater or lesser generality correspond to larger or smaller tiles, while services that adapt or develop services for novel tasks correspond to large tiles with greater initial latencies and task-specification requirements. The task-space model provides a conceptual, and perhaps practical, approach to implementing open-ended, asymptotically-comprehensive AI services as an effective form of general intelligence.

39.2 Broad capabilities call for mechanisms that compose diverse competencies

Current AI systems are notoriously narrow, and expanding the scope of functionality of individual systems is a major focus of research. Despite progress, vision networks that recognize faces are still distinct from networks that classify images, which are distinct from networks that parse scenes into regions corresponding to objects of different kinds—to say nothing of the differences between any of these and networks architected and trained to play Go or translate languages or predict the properties of molecules. Because it would be surprising to find that any single network architecture will be optimal for proposing Go moves, *and* identifying faces, *and* translating English to French, it is natural to ask how diverse, relatively narrow systems could be composed to form systems that externally present broad, seamless competencies.

From this perspective, matching tasks (inputs and goals) to services (*e.g.*, trained networks) is central to developing broadly-applicable intelligent functionality, whether the required task-matching mechanisms provide coherence to services developed within a system through the differentiation of subtasks, or integrate services developed independently. This perspective equally applicable to "hard-wiring" task-to-service matching at development time and dynamic matching during process execution, and equally applicable to deep-learning approaches and services implemented by (for example) Google's algorithm-agnostic AutoML. In practice, we should expect to see systems that exploit both static and dynamic matching, as well as specialized services implemented by relatively general service-development services (Li and Li 2018). Algorithm selection has long been an active field (Kotthoff 2012; Mısır and Sebag 2017; Yang et al. 2018).

39.3 The task-space concept suggests a model of integrated AI services

It is natural to think of services as populating task spaces in which similar services are neighbors and dissimilar services are distant, while broader services cover broader regions. This picture of services and task-spaces can be useful both as a *conceptual model* for thinking about broad AI competencies, and as a *potential mechanism* for implementing them.

1. *As a conceptual model*, viewing services as tiling a high-dimensional task space provides a framework for considering the relationship between tasks and services: In the task-space model, the diverse properties that

differentiate tasks are reflected in the high dimensionality of the task space, services of greater scope correspond to tiles of greater extent, and gaps between tiled regions correspond to services yet to be developed.

2. *As an implementation mechanism*, jointly embedding task and service representations in high dimensional vector spaces could potentially facilitate matching of tasks to services, both statically during implementation and dynamically during execution. While there is good reason to think that joint embedding will be a useful implementation technique, the value of task spaces as a conceptual model would stand even if alternative implementation techniques prove to be superior.

The discussion that follows will explore the role of vector embeddings in modern AI systems, first, to support proposition (1) by illustrating the richness and generality of vector representations, and, second, to support proposition (2) by illustrating the range of areas in which proximity-based operations on vector representations already play fundamental roles in AI system implementation. With a relaxed notion of "space", proposition (1) makes intuitive sense; the stronger proposition (2) requires closer examination.

39.4 Embeddings in high-dimensional spaces provide powerful representations

This discussion will assume a general familiarity with the unreasonable effectiveness of high-dimensional vector representations in deep learning systems, while outlining some relevant developments. In brief, deep learning systems often encode complex and subtle representations of the objects of a domain (be it an image, video, text, product) as numerical vectors ("embeddings") in spaces with tens to thousands of dimensions; the geometric relationships among embeddings encode relationships among the objects. In particular—given a successful embedding—similar objects (images of the same class, texts will similar meanings) map to vectors that are near neighbors in the embedding space.

Distances between vectors can be assigned in various ways: The most common in neural networks is cosine similarity, the inner product of vectors normalized to unit length; in high dimensional spaces, the cosine similarity between randomly-oriented vectors will, with high probability, be ≈ 0, while values substantially greater than 0 indicate relatively near neighbors. Distances in several other spaces have found use: The most common is Euclidian (L2) norm in conventional vector spaces, but more recent studies have employed Euclidian distance in toroidal spaces (placing bounds on distances

while preserving translational symmetry [Ebisu and Ichise 2017]), and distances in hyperbolic spaces mapped onto the Poincare ball (the hyperbolic metric is particularly suited to a range of graph embeddings: it allows volume to grow exponentially with distance, much as the width of a balanced tree grows exponentially with depth) (Gülçehre et al. 2018; Tifrea, Bécigneul, and Ganea 2018). Spaces with different metrics and topologies (and their Cartesian products) may be suited to different roles, and individually-continuous spaces can of course be disjoint, and perhaps situated in a discrete taxonomic space.

39.5 High-dimensional embeddings can represent semantically rich domains

In deep neural networks, every every vector that feeds into a fully-connected layer can be regarded as a vector-space embedding of some representation; examples include the top-level hidden layers in classifiers and intermediate encodings in encoder-decoder architectures. Applications of vector embedding have been extraordinarily broad, employing representations of:

- Images for classification tasks (Krizhevsky, Sutskever, and Hinton 2012)
- Images for captioning tasks (Vinyals et al. 2014)
- Sentences for translation tasks (Wu et al. 2016)
- Molecules for property-prediction tasks (Coley et al. 2017)
- Knowledge graphs for link-prediction tasks (Bordes et al. 2013)
- Products for e-commerce (J. Wang et al. 2018)

The successful application of vector embedding to diverse, semantically complex domains suggests that task-space models are not only coherent as a concept, but potentially useful in practice.

39.6 Proximity-based (application/activation/access) can deliver diverse services

Because applications may use neighborhood relationships to provide diverse functionality (here grouped under the elastic umbrella of "services"), the present discussion will refer to these collectively as "PBA operations", where "PB" denotes "proximity-based", and "A" can be interpreted as *application* of selected functions, *activation* of selected features, or more generically, *access* to (or retrieval of) selected entities. In each instance, PBA operations compute a measure of distance between a task representation embedding ("query") and a pre-computed embedding ("key") corresponding to the accessed feature,

Encoding and decoding vector embeddings

Image classification:

Image → *CNN* → **embedding** → *projection matrix* → class
(Krizhevsky, Sutskever, and Hinton 2012)

Image captioning:

Image → *CNN* → **embedding** → *RNN* → caption
(Vinyals et al. 2014)

Language translation:

Sentence → *RNN* → **embedding** → *RNN* → translation
(Wu et al. 2016)

function, or entity ("value"). In other words, PBA operations employ key/value lookup of near neighbors in a vector space. As shorthand, one can speak of values has having positions defined by their corresponding keys.

PBA operations may access multiple entities; when these are (or produce) embedding vectors, it can be useful to weight and add them (*e.g.*, weighting by cosine similarity between query and key). Weighted PBA (wPBA) operations that discard distant values are PBAs in a strict sense, and assigning small weights to distant entities has a similar effect. The class of wPBA operations thus includes any matrix multiplication in which input and row (= query and key) vectors are actually or approximately normalized (see Salimans and Kingma (2016) and C. Luo et al. (2017)), and the resulting activation vectors are sparse, *e.g.*, as a consequence of negatively biased ReLU units.

39.7 PBA operations are pervasive in deep learning systems

PBA mechanisms should not be regarded as a clumsy add-on to neural computation; indeed, the above example shows that wPBA operations can be found at the heart of multilayer perceptrons. A wide range of deep learning systems apply (what can be construed as) PBA operations to access (what can be construed as) fine-grained "services" *within* a neural network computation. For example, wPBA mechanisms have been used implement not only mixture-of-experts models (Shazeer et al. 2017; Kaiser, Gomez, et al. 2017), but also attention mechanisms responsible for wide-ranging advances in deep learning (Vaswani et al. 2017; Kool, van Hoof, and Welling 2018; Hudson and Manning 2018), including memories of past situations in RL agents (Wayne et al. 2018).

PBA mechanisms are prominent in single-shot learning in multiple domains (Kaiser, Nachum, et al. 2017), including learning new image classes from a single example. In the latter application, networks are trained to map images to an embedding space that supports classification; single-shot learning is performed by mapping pairs of novel labels and embeddings to that same embedding space, enabling subsequent classification by retrieving the label of the best-matching embedding ("best-matching" means closeness, *e.g.*, cosine similarity) (Vinyals et al. 2016). In one implementation, this is accomplished by normalizing and inserting the embedding of a new example in a standard projection matrix (Qi, Brown, and Lowe 2017) (indeed, with suitable normalization, standard image classification architectures can be regarded as employing PBA).

39.8 Joint embeddings can link related semantic domains

Embeddings can not only map similar entities to neighboring locations, but can also align distinct domains such that entities in one domain are mapped to locations near those of corresponding entities in the other (Frome et al. 2013; Y. Li et al. 2015; Baltrusaitis, Ahuja, and Morency 2017). Applications have been diverse:

- Text and images to enable image retrieval (K. Wang et al. 2016)
- Video and text to enable action recognition (Xu, Hospedales, and Gong 2017)
- Sounds and objects in video to learn cross-modal relationships (Arandjelović and Zisserman 2017)
- Images and recipes to retrieve one given the other (Salvador et al. 2017)
- Images and annotations to improve embeddings (Gong et al. 2014)
- Queries and factual statements to enable text-based question answering (Kumar et al. 2015)
- Articles and user-representations to recommend news stories (Okura et al. 2017)
- Product and user/query representations to recommend products (Zhang, Yao, and Sun 2017)

39.9 PBA operations can help match tasks to candidate services at scale

The breadth of applications noted above suggests that AI services and tasks could be represented and aligned in vector embedding spaces. This observa-

tion shows that the task-space concept is (at the very least!) coherent, but also suggests that PBA operations are strong candidates for actually implementing task-service matching. This proposition is compatible with the use of disjoint embedding spaces for different classes of tasks, the application of further selection criteria not well represented by distance metrics, and the use of PBA operations in developing systems that hard-wire services to sources of streams of similar tasks.

In considering joint embeddings of tasks and services, one should imagine joint training to align the representations of both service-requesting and service-providing components. Such representations would encode the nature of the task (vision? planning? language?), domain of application (scene? face? animal?), narrower domain specifications (urban scene? desert scene? Martian scene?), kind of output (object classes? semantic segmentation? depth map? warning signal?), and further conditions and constraints (large model or small? low or high latency? low or high resolution? web-browsing or safety-critical application?).

Note that exploitation of specialized services by diverse higher-level systems is in itself a form of transfer learning: To train a service for a task in one context is to train it for similar tasks wherever they may arise. Further, the ability to find services that are near-matches to a task can provide trained networks that are candidates for fine-tuning, or (as discussed below) architectures that are likely to be well-suited to the task at hand (Vanschoren 2018).

The concept of joint task-to-service embeddings suggests directions for experimental exploration: How could embeddings of training sets contribute to embeddings of trained networks? Distributional shifts will correspond to displacement vectors in task space—could regularities in those shifts be learned and exploited in metalearning? Could relationships among task embeddings guide architecture search? Note that task-to-service relationships form a bipartite graph in which links can be labeled with performance metrics; in optimized embeddings, the distances between tasks and services will be predictive of performance.

PBA operations can be applied at scale: A recently developed graph-based, polylogarithmic algorithm running can return sets of 100 near neighbors from sets of $>10^7$ vector embeddings with millisecond response times (single CPU) (Fu, Wang, and Cai 2017). Alibaba employs this algorithm for product recommendation at a scale of billions of items and customers (J. Wang et al. 2018).

39.10 PBA operations can help match new tasks to service-development services

Fluid matching of tasks to services tends to blunt the urgency of maximizing the scope of individual models. Although models with more general capabilities correspond to broader tiles in task-space, the size of individual tiles does not determine the breadth of their aggregate scope. Advances in metalearning push in the same direction: A task that maps to a gap between tiles may still fall within the scope of a reliable metalearning process that can, on demand, fill that gap (Vanschoren 2018). A particular metalearning system (characterized by both architecture and training) would in effect constitute a broad but high-latency tile which has first-use costs that include both data and computational resources for training. Graphs representing deep learning architectures can themselves be embedded in continuous spaces (and, remarkably, can be optimized by gradient descent [R. Luo et al. 2018]); learning and exploiting joint embeddings of tasks and untrained architectures would be a natural step.

In an intuitive spatial picture, metalearning methods enable population of a parallel space of service-providing services, a kind of backstop for tasks that pass through gaps between tiles in the primary task-space. Taking this picture further, one can picture a deeper backstop characterized by yet broader, more costly tiles: This space would be populated by AI research and development systems applicable to broader domains; such systems might search spaces of architectures, training algorithms, and data sets in order to provide systems suitable for filling gaps in metalearning and primary-task spaces. AI R&D comprises many subtasks (architecture recommendation, algorithm selection, *etc.*) that can again be situated in appropriate task spaces; as with other high-level services, we should expect high-level AI-development services to operate by delegating tasks and coordinating other, narrower services.

One may speculate that systems that display flexible, general intelligence will, internally, link tasks to capabilities by mechanisms broadly similar those in today's deep learning systems—which is to say, by mechanisms that can be construed as employing similarity of task and service embeddings in high-dimensional vector spaces. What is true of both multi-layer perceptrons and e-commerce recommendation systems is apt to be quite general.

39.11 Integrated, extensible AI services constitute general artificial intelligence

The concept of "general intelligence" calls for a capacity to learn and apply an indefinitely broad range of knowledge and capabilities, including high-level capabilities such as engineering design, scientific inquiry, and long-term planning. The concept of comprehensive AI services is the same: The CAIS model calls for the capacity to develop and apply an indefinitely broad range of services that provide both knowledge and capabilities, again including high-level services such as engineering design, scientific inquiry, and long-term planning. In other words, broad, extensible, integrated CAIS in itself constitutes general artificial intelligence, differing from the familiar AGI picture chiefly in terminology, concreteness, and avoidance of the long-standing assumption that well-integrated general intelligence necessarily entails unitary agency.

Further Reading

- *Section 1: R&D automation provides the most direct path to an intelligence explosion*
- *Section 12: AGI agents offer no compelling value*

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

> 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

Neurally-inspired AI systems implement a range of narrow yet recognizably human-like competencies, hence their computational costs and capabilities can provide evidence regarding the computational requirements of hypothetical systems that could deliver more general human-like competencies at human-like speeds. The present analysis relies on evidence linking task functionality to resource requirements in machines and biological systems, making no assumptions regarding the nature of neural structure or activity.

Comparisons in areas that include vision, speech recognition, and language translation suggest that affordable commercial systems (~1 PFLOP/s, costing $150,000 in 2017) may surpass brain-equivalent computational capacity,

perhaps by a substantial margin. Greater resources can be applied to learning, and the associated computational costs can be amortized across multiple performance-providing systems; current experience suggests that deep neural network (DNN) training can be fast by human standards, as measured by wall-clock time. In light of these considerations, it is reasonable to expect that, given suitable software, affordable systems will be able to perform human-level tasks at superhuman speeds.

40.2 Metrics and methodology

40.2.1 AI-technology performance metrics include both task *competencies* and task *throughput*

Hypothetical AI software that could perform tasks with human-level (or better) competence, in terms of scope and quality, would be constrained by contemporaneous computational capacity, and hence might perform with less-than-human task throughput; if so, then restricted hardware capacity might substantially blunt the practical implications of qualitative advances in AI software. By contrast, if advanced competencies were developed in the context of better-than-human hardware capacity ("hardware overhang"), then the practical implications of qualitative advances in AI software could potentially be much greater. A better understanding of the computational requirements for human-level performance (considering both competence and throughput) would enable a better understanding of AI prospects.

40.2.2 Ratios of hardware capacities and neural capacities (considered separately) compare apples to apples

The following analysis references an imperfect measure of real-world hardware capacity—floating-point performance—yet because it considers only *ratios* of capacity between broadly-similar systems applied to broadly-similar tasks, the analysis implicitly (though approximately) reflects cross-cutting considerations such as constraints on memory bandwidth. Thus, despite referencing an imperfect measure of hardware capacity, the present methodology compares apples to apples.

The neural side of the analysis is similar in this regard, considering only *ratios* of (estimates of) neural activity required for task performance relative to activity in the brain as a whole; these ratio-estimates are imperfect, but again compare apples to apples. Note that this approach avoids dubious comparisons of radically different phenomena such as synapse firing and

logic operations, or axonal signaling and digital data transmission. Neural structure and function is treated as a black box (as is AI software).

In the end, the quantity of interest (an estimate of machine capacity/brain capacity) will be expressed as a ratio of dimensionless ratios.

40.2.3 Laboratory-affordable AI hardware capacity reached ~1 PFLOP/s in 2017.

In 2017, NVIDIA introduced a 960 TFLOP/s "deep-learning supercomputer" (a 5.6× faster successor to their 2016 DGX-1 machine), at a price of

$150,000; a high-end 2017 supercomputer (Sunway TaihuLight) delivers ~100 PFLOP/s; a high-end 2017 gaming GPU delivers ~0.01 PFLOP/s. The following discussion will take 1 PFLOP/s as a reference value for current laboratory-affordable AI hardware capacity.

40.2.4 The computational cost of machine tasks scaled to human-like throughput is reasonably well defined

Consider machine tasks that are narrow in scope, yet human-comparable in quality: For a given machine task and implementation (*e.g.*, of image classification), one can combine a reported computational cost (in FLOP/s) and reported throughput (*e.g.*, frames per second) to define a cost scaled to human-like task-throughput (*e.g.*, image classification at 10 frames per second). Call this the "machine-task cost", which will be given as a fraction of a PFLOP/s.

40.2.5 Narrow AI tasks provide points of reference for linking computational costs to neural resource requirements

AI technologies based on neurally-inspired DNNs have achieved human-comparable capabilities on narrow tasks in domains that include vision, speech recognition, and language translation. It is difficult to formulate accurate, quantitative comparisons that link the known computational costs of narrow AI tasks to the resource costs of similar (yet never equivalent) neural tasks, yet for any given task comparison, one can encapsulate the relevant ambiguities and uncertainties in a single dimensionless parameter, and can consider the implications of alternative assumptions regarding its value.

A key concept in the following will be "immediate neural activity" (INA), an informal measure of *potentially task-applicable* brain activity. As a measure of current neural activity *potentially* applicable to task performance, INA is to

be interpreted in an abstract, information-processing sense that conceptually excludes the formation of long-term memories (as discussed below, human and machine learning are currently organized in fundamentally different ways).

The estimates of *task-applied* INA in this section employ cortical volumes that could be refined through closer study of the literature; a point of conservatism in these estimates is their neglect of the differential, task-focused patterns of neural activity that make fMRI informative (Heeger and Ress 2002). Differential activation of neural tissue for different tasks is analogous to the use of gated mixture-of-experts models in DNNs: In both cases, a managerial function selects and differentially activates task-relevant resources from a potentially much larger pool. In DNN applications (*e.g.*, language translation), a gated mixture-of-experts approach can increase model capacity by a factor of 100 to 1000 with little increase in computational cost (Shazeer et al. 2017).

40.2.6 The concept of a "task-INA fraction" encapsulates the key uncertainties and ambiguities inherent in linking machine-task costs to brain capacity

The present discussion employs the concept of a "task-INA fraction" (f_{INA}), the ratio between the INA that would (hypothetically) be required for a neural system to perform a given machine task and the contemporaneous global INA of a human brain (which may at a given moment support vision, motor function, auditory perception, higher-level cognition, *etc.*). This ratio encapsulates the main ambiguities and uncertainties in the chain of inference that links empirical machine performance to estimates of the requirements for human-equivalent computation. These ambiguities and uncertainties are substantial: Because no actual neural system performs the same task as a machine, any comparison of machine tasks to neural tasks can at best be approximate.

For example, convolutional neural networks (CNNs) closely parallel the human visual system in extracting image features, but the functional overlap between machine and neural tasks dwindles and disappears at higher levels of processing that, in CNNs, may terminate with object segmentation and classification. Potentially quantifiable differences between CNN and human visual processing include field of view, resolution, and effective frame rate. More difficult to disentangle or quantify, however, is the portion of visual-task INA that should be attributed to narrow CNN-like feature extraction, given that even low-level visual processing is intertwined with inputs that include feedback from higher levels, together with more general contextual

and attentional information (Heeger and Ress 2002).

Ambiguities and uncertainties of this kind increase when we consider tasks such as machine speech transcription (which is partially auditory and partially linguistic, but at a low semantic level), or language translation that is human-comparable in quality (Wu et al. 2016), yet employs a very limited representation of language-independent meaning (Johnson et al. 2016).

40.2.7 Uncertainties and ambiguities regarding values of $f_{\mathbf{INA}}$ are bounded

Despite these uncertainties and definitional ambiguities, there will always be bounds on plausible values of f_{INA} for various tasks. For, example, given that visual cortex occupies ~20% of the brain and devotes substantial resources to CNN-like aspects of feature extraction, it would be difficult to argue that the value of f_{INA} for CNN-like aspects of early visual processing is greater than 0.1 or less than 0.001. However, rather than emphasizing specific estimates of f_{INA} for specific machine tasks, the method of analysis adopted here invites the reader to consider the plausibility of a range of values based on some combination of knowledge from the neurosciences, introspection, and personal judgment. As we will see, even loose bounds on values of f_{INA} can support significant conclusions.

For a given task, a useful, empirically-based benchmark for comparison is the "PFLOP-parity INA faction" (f_{PFLOP}), which is simply the ratio of the empirical machine-task cost to a 1 PFLOP/s machine capacity. If the lowest plausible value of f_{INA} lies above the PFLOP-parity INA-faction for that same task, this suggests that a 1 PFLOP/s machine exceeds human capacity by a factor of $R_{\mathrm{PFLOP}} = \frac{f_{\mathrm{INA}}}{f_{\mathrm{PFLOP}}}$.

40.3 Estimated ratios for specific machine tasks

(Numbers in this section are rounded vigorously to avoid spurious implications of precision.)

40.3.1 Image-processing tasks vs. human vision tasks:

Systems based on Google's Inception architecture implement high-level feature extraction of a quality that supports comparable-to-human performance in discriminating among 1000 image classes. At a human-like 10 frames per second, the machine-task cost would be ~10 GFLOP/s (Szegedy et al. 2014), hence $f_{\mathrm{PFLOP}} = 10^{-5}$.

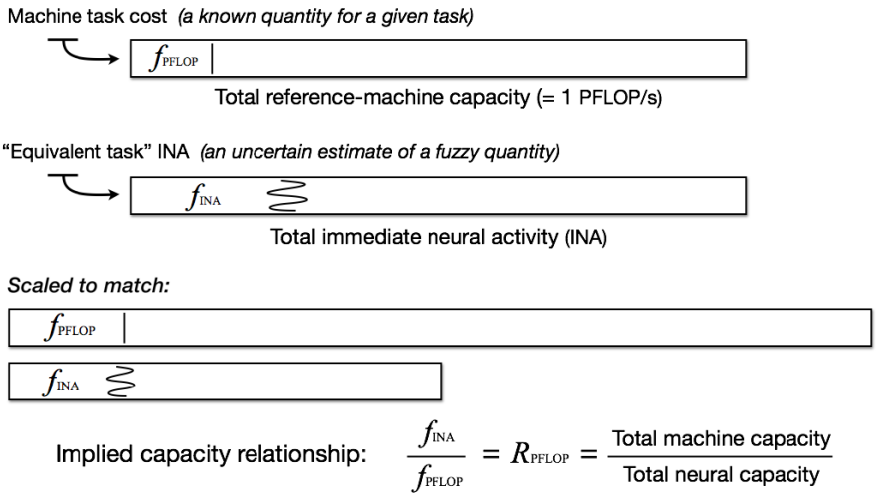

Figure 11: *Diagram of neural and computational costs as fractions of total resources (above), scaling the totals to align the fractions (below).*

Turning to neural function, consider that visual cortex comprises >10% of the human brain. If Inception-like high-level feature extraction were to require the equivalent of ~1% of visual cortex, then $f_ZINA = 10^{-3}$, and $R_{\text{PFLOP}} = 100$.

Speech-recognition tasks vs. human auditory/linguistic tasks: Baidu's Deep Speech 2 system can approach or exceed human accuracy in recognizing and transcribing spoken English and Mandarin, and would require approximately 1 GFLOP/s per real-time speech stream (Amodei et al. 2015). For this roughly human-level throughput, $f_{\text{PFLOP}} = 10^{-6}$.

Turning to neural function again, consider that task-relevant auditory/semantic cortex probably comprises >1% of the human brain. If the equivalent of the Deep Speech 2 speech-recognition task were to require 10% of that cortex, then $f_{\text{INA}} = 10^{-3}$, and $R_{\text{PFLOP}} = 1000$.

Language-translation tasks vs. human language comprehension tasks: Google's neural machine translation (NMT) systems have reportedly approached human quality (Wu et al. 2016). A multi-lingual version of the Google NMT model (which operates with the same resources) bridges language pairs through a seemingly language-independent representation of sentence meaning (Johnson et al. 2016), suggesting substantial (though unquantifiable) semantic depth in the intermediate processing. Performing

translation at a human-like rate of one sentence per second would require approximately 100 GFLOP/s, and $f_{\text{PFLOP}} = 10^{-4}$.

It is *plausible* that (to the extent that such things can be distinguished) human beings mobilize as much as 1% of global INA at an "NMT task-level"—involving vocabulary, syntax, and idiom, but not broader understanding—when performing language translation. If so, then for "NMT-equivalent translation," we can propose $f_{\text{INA}} = 10^{-2}$, implying $R_{\text{PFLOP}} = 100$.

Robotic vision vs. retinal visual processing: Hans Moravec applies a different yet methodologically similar analysis (Moravec 1998) that can serve as a cross-check on the above values. Moravec noted that both retinal visual processing and functionally-similar robot-vision programs are likely to be efficiently implemented in their respective media, enabling a comparison between the computational capacity of digital and neural systems. Taking computational requirements for retina-level robot vision as a baseline, then scaling from the volume of the retina to the volume of the brain, Moravec derives the equivalent of $R_{\text{PFLOP}} = \sim 10$ (if we take MIP/s ~ MFLOP/s). Thus, the estimates here overlap with Moravec's. In the brain, however, typical INA per unit volume is presumably less that that of activated retina, and a reasonable adjustment for this difference would suggest $R_{\text{PFLOP}} > 100$.

40.3.2 It seems likely that 1 PFLOP/s machines equal or exceed the human brain in raw computation capacity

In light of the above comparisons, all of which yield values of R_{PFLOP} in the 10 to 1000 range, it seems likely that 1 PFLOP/s machines equal or exceed the human brain in raw computation capacity. To draw the opposite conclusion would require that the equivalents of a *wide range* of seemingly substantial perceptual and cognitive tasks would *consistently* require no more than an *implausibly small fraction* of total neural activity.

The functional-capacity approach adopted here yields estimates that differ substantially, and sometimes greatly, from estimates based on proposed correspondences between neural activity and digital computation. Sandberg and Bostrom (Sandberg and Bostrom 2008), for example, consider brain emulation at several levels: analog neural network populations, spiking neural networks, and neural electrophysiology; the respective implied R_{PFLOP} values are 1, 10^{-3}, and 10^{-7}. Again based on a proposed neural-computational correspondence, Kurzweil suggests the equivalent of $R_{\text{PFLOP}} = 0.1$ (Kurzweil 2005).

40.4 Even with current methods, training can be fast by human standards

The discussion above addresses only task performance, but DNN technologies also invite a comparisons of machine and human learning speeds.

Human beings require months to years to learn to recognize objects, to recognize and transcribe speech, and to learn vocabulary and translate languages. Given abundant data and 1 PFLOP/s of processing power, the deep learning systems referenced above could be trained in hours (image and speech recognition, ~10 exaFLOPs) to weeks (translation, ~1000 exaFLOPs). These training times are short by human standards, which suggests that future learning algorithms running on 1 PFLOP/s systems could rapidly learn task domains of substantial scope. A recent systematic study shows that the scale of efficient parallelism in DNN training increases as tasks grow more complex, suggesting that training times could remain moderate even as product capabilities increase (McCandlish et al. 2018).

40.5 Large computational costs for training need not substantially undercut the implications of low costs for applications

Several considerations strengthen the practical implications of fast training, even if training for broad tasks were to require more extensive machine resources:

- More that 1 PFLOP/s can be applied to training for narrow AI tasks.
- Because broad capabilities can often be composed by coordinating narrower capabilities, parallel, loosely-coupled training processes may be effective in avoiding potential bottlenecks in learning broader AI tasks.
- In contrast to human learning, machine training costs can be amortized over an indefinitely large number of task-performing systems, hence training systems could be costly without undercutting the practical implications of high task-throughput with affordable hardware.

Human beings (unlike most current DNNs) can learn from single examples, and because algorithms with broad human-level competencies will (almost by definition) reflect solutions to this problem, we can expect the applicable training methods to be more efficient than those discussed above. Progress in "single-shot learning" is already substantial.

Note that hardware-oriented comparisons of speed do not address the qualitative shortcomings of current DNN training methods (*e.g.*, limited general-

ization, requirements for enormous amounts of training data). The discussion here addresses only quantitative measures (learning speed, task throughput).

40.6 Conclusions

Many modern AI tasks, although narrow, are comparable to narrow capacities of neural systems in the human brain. Given an empirical value for the fraction of computational resources required to perform that task with human-like throughput on a 1 PFLOP/s machine, and an inherently uncertain and ambiguous—yet bounded—estimate of the fraction of brain resources required to perform "the equivalent" of that machine task, we can estimate the ratio of PFLOP/s machine capacity to brain capacity. What are in the author's judgment plausible estimates for each task are consistent in suggesting that this ratio is ~10 or more. Machine learning and human learning differ in their relationship to costs, but even large machine learning costs can be amortized over an indefinitely large number of task-performing systems and application events.

In light of these considerations, we should expect that substantially super-human computational capacity will accompany the eventual emergence of a software with broad functional competencies. On present evidence, scenarios that assume otherwise seem unlikely.

Further Reading

- *Section 12: AGI agents offer no compelling value*
- *Section 23: AI development systems can support effective human guidance*
- *Section 24: Human oversight need not impede fast, recursive AI technology improvement*
- *Section 36: Desiderata and directions for interim AI safety guidelines*

Afterword

While this document was being written, AI researchers have, as the R&D-automation/AI-services model would predict, continued to automate research and development processes while developing systems that apply increasingly general learning capabilities to an increasing range of tasks in bounded domains. Progress along these lines continues to exceed my expectations in surprising ways, particularly in the automation of architecture search and training.

The central concepts presented in this document are intended to be what Chip Morningstar calls "the second kind of obvious"—obvious once pointed out, which is to say, obvious in light of facts that are already well-known. Based on the reception in the AI research community to date, this effort seems to have largely succeeded.

Looking forward, I hope to see the comprehensive AI-services model of general, superintelligent-level AI merge into the background of assumptions that shape thinking about the trajectory of AI technology. Whatever one's expectations may be regarding the eventual development of advanced, increasingly general AI agents, we should expect to see diverse, increasingly general superintelligent-level services as their predecessors and as components of a competitive world context. This is, I think, a robust conclusion that reframes many concerns.

Acknowledgements

For questions, ideas, perspectives, criticisms ranging from profound to stylistic, the author thanks (with apologies to those omitted) Dario Amodei, Shahar Avin, Nick Beckstead, Devi Borg, Nick Bostrom, Miles Brundage, Ryan Carey, Joseph Carlsmith, Paul Christiano, Owen Cotton-Barratt, Andrew Critch, Chris Cundy, Allan Dafoe, Daniel Dewey, Jeffrey Ding, Owain Evans, Aleš Flídr, Carrick Flynn, Victoria Krakovna, Dave Krueger, Shane Legg, Jan Leike, Jade Leung, Jelena Luketina, Laurent Orseau, Scott Garrabrant, Ben Garfinkel, Zac Kenton, Adam Marblestone, Tom McGrath, Mark S. Miller, Luke Muehlhauser, Richard Ngo, Seán Ó hÉigeartaigh, Toby Ord, Laurent Orseau, Pedro Ortega, Mike Page, Stuart Russell, Anders Sandberg, Kyle Scott, Rohin Shah, Andrew Snyder-Beattie, Nate Soares, Jaan Tallinn, Jessica Taylor, Eric Tribble, Eliezer Yudkowsky, and Rosa Wang, as well as audiences at DeepMind, OpenAI, Berkeley's Center for Human-Compatible AI, and the Machine Intelligence Research Institute, among others.

I also thank Tanya Singh for launching the LaTeX conversion project, Justis Mills for breaking ground and setting direction, Jimmy Rintjema for completing the project with speed and flair, and the Berkeley Existential Risk Initiative (BERI) for paying the bills.

Special thanks are due to Nick Bostrom for formulating hard questions and providing an environment in which they could be studied, and to the chairman of my doctoral committee, Marvin Minsky, for encouraging me, in a conversation *circa* 1990, to write up a game-theoretic approach for obtaining trustworthy answers from untrusted superintelligent-level AI systems; although I procrastinated for over 25 years, the approach has now been refined and presented in Section 16, "Collusion among superintelligent oracles can readily be avoided", following related discussions in Drexler (2015).

This work was supported by grants from the European Research Council and the Alexander Tamas Programme on AI Safety, and made possible by the Future of Humanity Institute at the University of Oxford.

The cover art is excerpted from a 2005 map of the Internet (credit: Wikimedia Commons)

References

Abolafia, Daniel A., Mohammad Norouzi, Jonathan Shen, Rui Zhao, and Quoc V. Le. 2018. "Neural Program Synthesis with Priority Queue Training." arXiv: 1801.03526 [cs.AI].

Allamanis, Miltiadis, and Marc Brockschmidt. 2017. "SmartPaste: Learning to Adapt Source Code." arXiv: 1705.07867 [cs.LG].

Allamanis, Miltiadis, Marc Brockschmidt, and Mahmoud Khademi. 2018. "Learning to Represent Programs with Graphs." arXiv: 1711.00740 [cs.LG].

Amodei, Dario, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, et al. 2015. "Deep speech 2: End-to-end speech recognition in English and Mandarin." arXiv: 1512.02595 [cs.CL].

Amos, Brandon, and J. Zico Kolter. 2017. "Optnet: Differentiable optimization as a layer in neural networks." arXiv: 1703.00443 [cs.LG].

Andreas, Jacob, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. "Neural Module Networks." In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anthony, Thomas, Zheng Tian, and David Barber. 2017. "Thinking Fast and Slow with Deep Learning and Tree Search." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5360–5370. Curran Associates, Inc.

Arandjelović, Relja, and Andrew Zisserman. 2017. "Objects that sound." *arXiv preprint arXiv:1712.06651*.

Armstrong, Stuart. 2013. "Domesticating reduced impact AIs." *Less Wrong* (blog). https://www.lesswrong.com/posts/FdcxknHjeNH2MzrTj/domesticating-reduced-impact-ais.

Armstrong, Stuart, and Benjamin Levinstein. 2017. "Low Impact Artificial Intelligences." arXiv: 1705.10720 [cs.AI]. http://arxiv.org/abs/1705.10720.

Baker, Bowen, Otkrist Gupta, Nikhil Naik, and Ramesh Raskar. 2016. "Designing Neural Network Architectures using Reinforcement Learning." arXiv: 1611.02167 [cs.LG].

Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. "Multimodal Machine Learning: A Survey and Taxonomy." arXiv: 1705.09406 [cs.LG].

Battaglia, Peter, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. 2016. "Interaction Networks for Learning about Objects, Relations and Physics." In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 4502–4510. Curran Associates, Inc.

Bello, Irwan, Barret Zoph, Vijay Vasudevan, and Quoc V. Le. 2017. "Neural Optimizer Search with Reinforcement Learning." arXiv: 1709.07417 [cs.LG].

Besold, Tarek R., Artur S. d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro M. Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, et al. 2017. "Neural-Symbolic Learning and Reasoning: A Survey and Interpretation." arXiv: 1711.03902 [cs.AI].

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. "Translating Embeddings for Modeling Multi-relational Data." In *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, 2787–2795. Curran Associates, Inc.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* New York: Oxford University Press.

Britz, Denny, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. "Massive Exploration of Neural Machine Translation Architectures." arXiv: 1703.03906 [cs.CL].

Burda, Yuri, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. 2018. "Large-scale study of curiosity-driven learning." *arXiv preprint arXiv:1808.04355.*

Carbune, Victor, Thierry Coppey, Alexander Daryin, Thomas Deselaers, Nikhil Sarda, and Jay Yagnik. 2017. "Predicted Variables in Programming." arXiv: 1810.00619 [cs.LG].

Chandar, Sarath, Sungjin Ahn, Hugo Larochelle, Pascal Vincent, Gerald Tesauro, and Yoshua Bengio. 2016. "Hierarchical Memory Networks." arXiv: 1605.07427 [stat.ML].

Chang, Michael B., Tomer Ullman, Antonio Torralba, and Joshua B. Tenenbaum. 2016. "A Compositional Object-Based Approach to Learning Physical Dynamics." arXiv: 1612.00341 [cs.AI].

Chen, Yutian, Matthew W. Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P. Lillicrap, Matt Botvinick, and Nando de Freitas. 2017. "Learning to Learn without Gradient Descent by Gradient Descent." In *Proceedings of the 34th International Conference on Machine Learning*, edited by Doina Precup and Yee Whye Teh, 70:748–756. Proceedings of Machine Learning Research.

Chollet, François. 2017. "The impossibility of intelligence explosion." *Medium* (medium.com). https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec.

Christiano, Paul. 2014. "Approval-directed agents." *AI Alignment* (medium.com). https://ai-alignment.com/model-free-decisions-6e6609f5d99e.

———. 2015a. "On heterogeneous objectives." *AI Alignment* (medium.com). https://ai-alignment.com/on-heterogeneous-objectives-b38d0e003399.

———. 2015b. "Research directions in AI control." *AI Control* (medium.com). https://medium.com/ai-control/research-directions-in-ai-control-ef6f666d2062.

Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep reinforcement learning from human preferences." arXiv: 1706.03741 [stat.ML].

Clark, Jack. 2015. "Google Turning Its Lucrative Web Search Over to AI Machines." *Bloomberg*. https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines.

Coley, Connor W., Regina Barzilay, William H. Green, Tommi S. Jaakkola, and Klavs F. Jensen. 2017. "Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction." *Journal of Chemical Information and Modeling* 57 (8): 1757–1772.

Conti, Edoardo, Vashisht Madhavan, Felipe Petroski Such, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. 2017. "Improving Exploration in Evolution Strategies for Deep Reinforcement Learning via a Population of Novelty-Seeking Agents." arXiv: 1712.06560 [cs.AI].

Denil, Misha, Sergio Gomez Colmenarejo, Serkan Cabi, David Saxton, and Nando de Freitas. 2017. "Programmable Agents." arXiv: 1706.06383 [cs.AI].

Drexler, K. Eric. 2015. *MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities.* Technical Report 2015-3. Future of Humanity Institute, Oxford University. http://www.fhi.ox.ac.uk/wp-content/uploads/MDL-Intelligence-Distillation-for-safe-superintelligent-problem-solving1.pdf.

Duan, Yan, Marcin Andrychowicz, Bradly Stadie, OpenAI Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. 2017. "One-shot imitation learning." In *Advances in neural information processing systems*, 1087–1098.

Duan, Yan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. "RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning." arXiv: 1611.02779 [cs.AI].

Dyer, Chris, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. "Transition-based dependency parsing with stack long short-term memory." arXiv: 1505.08075 [cs.CL].

Ebisu, Takuma, and Ryutaro Ichise. 2017. "TorusE: Knowledge Graph Embedding on a Lie Group." arXiv: 1711.05435 [cs.AI].

Ehrhardt, Sébastien, Aron Monszpart, Niloy J. Mitra, and Andrea Vedaldi. 2017. "Learning a physical long-term predictor." arXiv: 1703.00247 [cs.AI].

Esmaeilzadeh, Hadi, Adrian Sampson, Luis Ceze, and Doug Burger. 2012. "Neural Acceleration for General-Purpose Approximate Programs." In *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, 449–460. MICRO-45. Vancouver, B.C., CANADA: IEEE Computer Society.

Evans, Richard, and Edward Grefenstette. 2018. "Learning Explanatory Rules from Noisy Data." *Journal of Artificial Intelligence Research* 61:1–64.

Farquhar, Gregory, Tim Rocktäschel, Maximilian Igl, and Shimon Whiteson. 2017. "TreeQN and ATreeC: Differentiable Tree Planning for Deep Reinforcement Learning." arXiv: 1710.11417 [cs.AI].

Frome, Andrea, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. "DeViSE: A Deep Visual-Semantic Embedding Model." In *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, 2121–2129. Curran Associates, Inc.

Fu, Cong, Changxu Wang, and Deng Cai. 2017. "Fast Approximate Nearest Neighbor Search With Navigating Spreading-out Graphs." arXiv: 1707.00143 [cs.LG].

Garcia, Victor, and Joan Bruna. 2017. "Few-Shot Learning with Graph Neural Networks." arXiv: 1711.04043 [stat.ML].

Garnelo, Marta, Kai Arulkumaran, and Murray Shanahan. 2016. "Towards Deep Symbolic Reinforcement Learning." arXiv: 1609.05518 [cs.AI].

George, Dileep, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan, Bhaskara Marthi, Xinghua Lou, et al. 2017. "A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs." *Science* 358 (6368).

Gilmer, Justin, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. "Neural message passing for quantum chemistry." arXiv: 1704.01212 [cs.LG].

Gong, Yunchao, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. "Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections." In *Computer Vision – ECCV 2014*, edited by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, 529–545. Cham: Springer International Publishing.

Good, Irving John. 1966. "Speculations Concerning the First Ultraintelligent Machine." *Advances in Computers* 6.

Goudet, Olivier, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. 2017. "Causal Generative Neural Networks." arXiv: 1711.08936 [stat.ML].

Graves, Alex, Greg Wayne, and Ivo Danihelka. 2014. "Neural turing machines." arXiv: 1410.5401 [cs.NE].

Graves, Alex, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, et al. 2016. "Hybrid computing using a neural network with dynamic external memory." *Nature* 538:471–476.

Grefenstette, Edward, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. 2015. "Learning to Transduce with Unbounded Memory." In *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, 1828–1836. Curran Associates, Inc.

Guez, Arthur, Théophane Weber, Ioannis Antonoglou, Karen Simonyan, Oriol Vinyals, Daan Wierstra, Rémi Munos, and David Silver. 2018. "Learning to Search with MCTSnets." arXiv: 1802.04697 [cs.AI].

Gülçehre, Çaglar, Sarath Chandar, and Yoshua Bengio. 2017. "Memory Augmented Neural Networks with Wormhole Connections." arXiv: 1701.08718 [cs.LG].

Gülçehre, Çaglar, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, et al. 2018. "Hyperbolic Attention Networks." arXiv: 1805.09786 [cs.NE].

Ha, David, and Jürgen Schmidhuber. 2018. "World Models." *arXiv preprint arXiv:1803.10122.*

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. "Cooperative Inverse Reinforcement Learning." arXiv: 1606.03137 [cs.AI].

Hanson, Robin. 2016. *The Age of Em: Work, Love, and Life when Robots Rule the Earth.* Oxford University Press.

Heeger, David J., and David Ress. 2002. "What does fMRI tell us about neuronal activity?" *Nature Reviews Neuroscience* 3:142–151.

Heess, Nicolas, Dhruva TB, Srinivasan Sriram, Jay Lemmon, Josh Merel, Greg Wayne, Yuval Tassa, et al. 2017. "Emergence of Locomotion Behaviours in Rich Environments." arXiv: 1707.02286 [cs.AI]. http://arxiv.org/abs/1707.02286.

Hu, Ronghang, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. "Explainable Neural Computation via Stack Neural Module Networks." arXiv: 1807.08556 [cs.CV].

Hudson, Drew A., and Christopher D. Manning. 2018. "Compositional Attention Networks for Machine Reasoning." arXiv: 1803.03067 [cs.AI].

Irving, Geoffrey, Christian Szegedy, Alexander A. Alemi, Niklas Een, Francois Chollet, and Josef Urban. 2016. "DeepMath - Deep Sequence Models for Premise Selection." In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 2235–2243. Curran Associates, Inc.

Jaderberg, Max, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, et al. 2017. "Population Based Training of Neural Networks." arXiv: 1711.09846 [cs.LG].

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, et al. 2016. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." arXiv: 1611.04558 [cs.CL].

Kaiser, Łukasz, Aidan N. Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. 2017. "One Model To Learn Them All." arXiv: 1706.05137 [cs.LG].

Kaiser, Łukasz, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. "Learning to Remember Rare Events." arXiv: 1703.03129 [cs.LG].

Kaiser, Łukasz, and Ilya Sutskever. 2015. "Neural GPUs learn algorithms." arXiv: 1511.08228 [cs.LG].

Kaliszyk, Cezary, François Chollet, and Christian Szegedy. 2017. "Holstep: A Machine Learning Dataset for Higher-Order Logic Theorem Proving." arXiv: 1703.00426 [cs.AI].

Kant, Neel. 2018. "Recent Advances in Neural Program Synthesis." arXiv: 1802.02353 [cs.AI].

Kelly, Kevin. 2017. "The Myth of a Superhuman AI." *Wired* (wired.com). https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/.

Klein, Gerwin, June Andronick, Kevin Elphinstone, Toby Murray, Thomas Sewell, Rafal Kolanski, and Gernot Heiser. 2014. "Comprehensive Formal Verification of an OS Microkernel." *ACM Transactions on Computer Systems* (New York, NY, USA) 32 (1): 2:1–2:70.

Kool, Wouter, Herke van Hoof, and Max Welling. 2018. "Attention Solves Your TSP, Approximately." arXiv: 1803.08475 [stat.ML].

Kotseruba, Iuliia, Oscar J. Avella Gonzalez, and John K. Tsotsos. 2016. "A Review of 40 Years of Cognitive Architecture Research: Focus on Perception, Attention, Learning and Applications." arXiv: 1610.08602 [cs.AI].

Kotthoff, Lars. 2012. "Algorithm Selection for Combinatorial Search Problems: A Survey." arXiv: 1210.7959 [cs.AI].

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105. Curran Associates, Inc.

Kumar, Ankit, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. "Ask Me Anything: Dynamic Memory Networks for Natural Language Processing." arXiv: 1506.07285 [cs.CL].

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Li, Fei-Fei, and Jia Li. 2018. "Cloud AutoML: Making AI accessible to every business." *The Keyword* (blog). https://ai.googleblog.com/2017/05/using-machine-learning-to-explore.html.

Li, Jian, Yue Wang, Irwin King, and Michael R. Lyu. 2017. "Code Completion with Neural Attention and Pointer Networks." arXiv: 1711.09573 [cs.CL].

Li, Jun, Kai Xu, Siddhartha Chaudhuri, Ersin Yumer, Hao Zhang, and Leonidas Guibas. 2017. "GRASS: Generative Recursive Autoencoders for Shape Structures." *ACM Transactions on Graphics* (New York, NY, USA) 36 (4): 52:1–52:14.

Li, Shen, Hengru Xu, and Zhengdong Lu. 2018. "Generalize Symbolic Knowledge With Neural Rule Engine." arXiv: 1808.10326 [cs.CL].

Li, Yangyan, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J. Guibas. 2015. "Joint Embeddings of Shapes and Images via CNN Image Purification." *ACM Transactions on Graphics* (New York, NY) 34 (6): 234:1–234:12.

Liao, Qianli, and Tomaso Poggio. 2017. *Object-Oriented Deep Learning.* Technical report, CBMM Memo Series 070. MIT Center for Brains, Minds and Machines (CBMM). http://hdl.handle.net/1721.1/112103.

Luo, Chunjie, Jianfeng Zhan, Lei Wang, and Qiang Yang. 2017. "Cosine Normalization: Using Cosine Similarity Instead of Dot Productin Neural Networks." arXiv: 1702.05870 [cs.LG].

Luo, Renqian, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2018. "Neural Architecture Optimization." arXiv: 1808.07233 [cs.LG].

McCandlish, Sam, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. 2018. "An Empirical Model of Large-Batch Training." *arXiv preprint arXiv:1812.06162.*

Merk, Daniel, Lukas Friedrich, Francesca Grisoni, and Gisbert Schneider. 2018. "De Novo Design of Bioactive Small Molecules by Artificial Intelligence." *Molecular Informatics* 37 (1-2): 1700153.

Mısır, Mustafa, and Michèle Sebag. 2017. "Alors: An algorithm recommender system." Combining Constraint Solving with Mining and Learning, *Artificial Intelligence* 244:291–314.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2015. "Human-level control through deep reinforcement learning." *Nature* 518:529–533.

Moravec, Hans P. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1.

Nair, Ashvin V, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. 2018. "Visual reinforcement learning with imagined goals." In *Advances in Neural Information Processing Systems*, 9208–9219.

Neelakantan, Arvind, Quoc V. Le, and Ilya Sutskever. 2015. "Neural Programmer: Inducing Latent Programs with Gradient Descent." arXiv: 1511.04834 [cs.LG].

Okura, Shumpei, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. "Embedding-based News Recommendation for Millions of Users." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1933–1942. KDD '17. Halifax, NS, Canada: ACM.

Pathak, Deepak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. "Curiosity-driven exploration by self-supervised prediction." In *International Conference on Machine Learning (ICML)*, vol. 2017.

Peng, Xue Bin, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. 2018. "SFV: reinforcement learning of physical skills from videos." In *SIGGRAPH Asia 2018 Technical Papers*, 178. ACM.

Pham, Hieu, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. 2018. "Efficient Neural Architecture Search via Parameter Sharing." arXiv: 1802.03268 [cs.LG].

Qi, Hang, Matthew Brown, and David G. Lowe. 2017. "Learning with Imprinted Weights." arXiv: 1712.07136 [cs.CV].

Raiman, Jonathan, and Olivier Raiman. 2018. "DeepType: Multilingual Entity Linking by Neural Type System Evolution." arXiv: 1802.01021 [cs.CL].

Sabour, Sara, Nicholas Frosst, and Geoffrey E Hinton. 2017. "Dynamic Routing Between Capsules." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 3856–3866. Curran Associates, Inc.

Salimans, Tim, and Diederik P. Kingma. 2016. "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks." arXiv: 1602.07868 [cs.LG].

Salvador, Amaia, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. "Learning Cross-modal Embeddings for Cooking Recipes and Food Images." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/wp-content/uploads/brain-emulation-roadmap-report1.pdf.

Schrimpf, Martin, Stephen Merity, James Bradbury, and Richard Socher. 2017. "A Flexible Approach to Automated RNN Architecture Generation." arXiv: 1712.07316 [cs.CL].

Sculley, D., Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. 2015. "Hidden Technical Debt in Machine Learning Systems." In *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, 2503–2511. Curran Associates, Inc.

Segler, Marwin H. S., Mike Preuß, and Mark P. Waller. 2017. "Towards "alphachem": Chemical synthesis planning with tree search and deep neural network policies." arXiv: 1702.00020 [cs.AI].

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." arXiv: 1701.06538 [cs.CL].

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529:484–489.

Singh, Rishabh, and Pushmeet Kohli. 2017. "AP: Artificial Programming." In *2nd Summit on Advances in Programming Languages (SNAPL 2017)*, edited by Benjamin S. Lerner, Rastislav Bodík, and Shriram Krishnamurthi, 71:16:1–16:12. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Snell, Jake, Kevin Swersky, and Richard Zemel. 2017. "Prototypical Networks for Few-shot Learning." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4077–4087. Curran Associates, Inc.

Soares, Nate. 2018. "The Value Learning Problem." Chap. 7 in *Artificial Intelligence Safety and Security*, edited by Roman Yampolskiy. Boca Raton, FL: CRC Press.

Steger, Carsten, Markus Ulrich, and Christian Wiedemann. 2007. *Machine Vision Algorithms and Applications.* John Wiley & Sons.

Sukhbaatar, Sainbayar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. "End-To-End Memory Networks." In *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, 2440–2448. Curran Associates, Inc.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. "Going Deeper with Convolutions." arXiv: 1409.4842 [cs.CV].

Tatarchenko, Maxim, Alexey Dosovitskiy, and Thomas Brox. 2017. "Octree Generating Networks: Efficient Convolutional Architectures for High-Resolution 3D Outputs." arXiv: 1703.09438 [cs.CV].

Teh, Yee Whye, Victor Bapst, Wojciech Marian Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. 2017. "Distral: Robust Multitask Reinforcement Learning." arXiv: 1707.04175 [cs.LG]. http://arxiv.org/abs/1707.04175.

Tifrea, Alexandru, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. "Poincaré GloVe: Hyperbolic Word Embeddings." arXiv: 1810.06546 [cs.CL].

Turing, Alan M. 1950. "Computing machinery and intelligence." *Mind* 59 (236): 433.

van den Oord, Aaron, Oriol Vinyals, and Koray Kavukcuoglu. 2017. "Neural Discrete Representation Learning." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 6306–6315. Curran Associates, Inc.

Vanschoren, Joaquin. 2018. "Meta-Learning: A Survey." arXiv: 1810.03548 [cs.LG].

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv: 1706.03762 [cs.CL].

Vinyals, Oriol, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. "Matching Networks for One Shot Learning." In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 3630–3638. Curran Associates, Inc.

Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. 2015. "Pointer Networks." In *Advances in Neural Information Processing Systems 28*, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, 2692–2700. Curran Associates, Inc.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. "Show and Tell: A Neural Image Caption Generator." arXiv: 1411.4555 [cs.CV].

Wang, Jane X., Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Rémi Munos, Charles Blundell, Dharshan Kumaran, and Matthew Botvinick. 2016. "Learning to reinforcement learn." arXiv: 1611.05763 [cs.LG].

Wang, Jizhe, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. 2018. "Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba." arXiv: 1803.02349 [cs.IR]. http://arxiv.org/abs/1803.02349.

Wang, Kaiye, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. 2016. "Joint Feature Selection and Subspace Learning for Cross-Modal Retrieval." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (10): 2010–2023.

Watters, Nicholas, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. 2017. "Visual Interaction Networks: Learning a Physics Simulator from Video." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4539–4547. Curran Associates, Inc.

Wayne, Greg, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack W. Rae, et al. 2018. "Unsupervised Predictive Memory in a Goal-Directed Agent." arXiv: 1803.10760 [cs.LG].

Weber, Théophane, Sébastien Racanière, David P Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adria Puigdomenech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. 2017. "Imagination-augmented agents for deep reinforcement learning." *arXiv preprint arXiv:1707.06203*.

Wolchover, Natalie. 2015. "Concerns of an Artificial Intelligence Pioneer." *Quanta Magazine* (quantamagazine.org). https://www.quantamagazine.org/artificial-intelligence-aligned-with-human-values-qa-with-stuart-russell-20150421.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. 2016. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." arXiv: 1609.08144 [cs.CL].

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." In *Proceedings of the 32nd International Conference on Machine Learning*, edited by Francis Bach and David Blei, 37:2048–2057. Proceedings of Machine Learning Research. PMLR.

Xu, Xun, Timothy Hospedales, and Shaogang Gong. 2017. "Transductive Zero-Shot Action Recognition by Word-Vector Embedding." *International Journal of Computer Vision* 123 (3): 309–333.

Yang, Chengrun, Yuji Akimoto, Dae Won Kim, and Madeleine Udell. 2018. "OBOE: Collaborative Filtering for AutoML Initialization." arXiv: 1808.03233 [cs.LG].

Yi, Kexin, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." arXiv: 1810.02338 [cs.AI].

Yin, Pengcheng, and Graham Neubig. 2017. "A syntactic neural model for general-purpose code generation." arXiv: 1704.01696 [cs.CL].

Yoon, KiJung, Renjie Liao, Yuwen Xiong, Lisa Zhang, Ethan Fetaya, Raquel Urtasun, Richard Zemel, and Xaq Pitkow. 2018. "Inference in Probabilistic Graphical Models by Graph Neural Networks." arXiv: 1803.07710 [cs.LG].

Zhang, Shuai, Lina Yao, and Aixin Sun. 2017. "Deep Learning based Recommender System: A Survey and New Perspectives." arXiv: 1707.07435 [cs.IR].

Zhou, Li, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot." arXiv: 1812.08989 [cs.HC].

Zoph, Barrett, and Quoc V. Le. 2016. "Neural architecture search with reinforcement learning." arXiv: 1611.01578 [cs.LG].

Zoph, Barrett, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2017. "Learning Transferable Architectures for Scalable Image Recognition." arXiv: 1707.07012 [cs.LG].