FHI TECHNICAL REPORT

MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities

K. Eric Drexler

Technical Report #2015-3

Cite as:

Drexler, K.E. (2015): "MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities", *Technical Report* #2015-3, Future of Humanity Institute, Oxford University: pp. 1-17.

MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities

K. Eric Drexler

April 2015

Overview

AI technologies may reach the threshold of rapid, open-ended, recursive improvement before we are prepared to manage the challenges posed by the emergence of superintelligent AI agents.[1] If this situation occurs, then it may become critically important to employ methods for reducing AI risks until more comprehensive solutions are both understood and ready for implementation. If methods for risk reduction can contribute to those comprehensive solutions, so much the better.

A foundational technique for reducing AI risks would apply capabilities for recursive AI improvement to a particular task: a process of "intelligence distillation" in which the metric for AI improvement is minimization of the description length of implementations that are themselves capable of open-ended recursive improvement.

By separating knowledge from learning capability, intelligence distillation can support strategies for implementing specialised, low-risk, yet superintelligent problem-solvers: Distillation can constrain initial information content; knowledge metering can constrain the information input during learning; checkpoint/restart protocols can constrain the retention of information provided in conjunction with tasks. Building on these methods and their functional products, sets of problem-solvers with superintelligent domain competencies could potentially be combined to implement highly capable systems that lack characteristics necessary for strong, risky AI agency. An appendix outlines how this strategy might be applied to implement superintelligent, human-interactive engineering systems with minimal risk.

Distillation/specialisation/composition strategies raise wide-ranging questions regarding the potential scope of safe applications of superintelligence-enabled AI capabilities. Because distillation-enabled strategies may offer practical means for mitigating AI risks while pursuing ambitious applications, further studies in this area could strengthen links between the AI-development and AI-safety research communities.

---

[1] Nick Bostrom's recent book, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), provides the broadest and deepest exploration of these challenges to date; the present document is intended for an audience that has a general familiarity with the considerations and problems addressed in *Superintelligence.*

1 Transitional AI safety: addressing the difficult case

In *Superintelligence* (Oxford University Press, 2014), Nick Bostrom explores a range of profound problems posed by the potential emergence of superintelligent AI agency, and suggests that adequate solutions may be long delayed. If AI technologies reach the threshold of rapid, open-ended, recursive improvement before we have full solutions to the problems explored in *Superintelligence*, then interim strategies for shaping and managing emerging superintelligence could be crucial.

In the reference problem-situation assumed here:

1) AI technology has reached the threshold of rapid, open-ended, recursive improvement.
2) The content and mechanisms of emerging superintelligent systems are effectively opaque,
3) Ongoing pressures for AI applications ensure that superintelligence will be widely exploited, and
4) No fully adequate solution to the problems posed by superintelligent agency is ready for implementation.

Conditions (1) through (4) are challenging, yet they are compatible with potentially powerful and accessible risk-reduction strategies. (These strategies could of course be applied under less challenging circumstances.)

In considering the force of point (3), one must keep in mind the ongoing pressures to apply advanced AI capabilities, including the sheer momentum of competitive research and development. Applications of superintelligence could not only be extraordinarily profitable, but could greatly augment scientific knowledge, global material wealth, human health, and perhaps even genuine security. Because it would be unwise to assume that emerging superintelligence will not be applied, there is good reason to seek means for implementing low-risk applications.

Disclaimer: After a talk on this topic at the Future of Humanity Institute on 4 Dec 2014, Anders Sandberg suggested that I write a brief summary, but although this document follows the content of the talk, it neither minimizes the description length of the concepts, nor adds the apparatus of scholarly citation.

Historical note: My concerns regarding AI risk, which center on the challenges of long-term AI governance, date from the inception of my studies of advanced molecular technologies, *ca.* 1977. I recall a later conversation with Marvin Minsky (then chairing my doctoral committee, *ca.* 1990) that sharpened my understanding of some of the crucial considerations: Regarding goal hierarchies, Marvin remarked that the high-level task of learning language is, for an infant, a *sub*goal of getting a drink of water, and that converting the resources of the universe into computers is a potential subgoal of a machine attempting to play perfect chess. The ideas presented here emerged as subgoals of proposed strategies for managing untrustworthy AI systems that I outlined to Marvin around the same time. He suggested that I do a write up; procrastination ensued.

From a risk-reduction perspective, transitional AI safety measures offer several potential benefits:

1) They can extend the time available for studying the fundamental problems of long-term AI control;
2) They can enable experimentation with operational and potentially surprising AI technologies; and, perhaps crucially,
3) They may enable the application of superintelligent problem-solving capabilities to the problem of managing superintelligence.

1.1 High and low-risk AI paths compared

Table 1 contrasts a potential AI development path that leads to severe AI-agent risk with a proposed path that would develop and apply superintelligent capabilities by means that could potentially obviate these risks.

Table 1. Potential paths to unsafe AI agents *vs.* low-risk AI tools:

A potential path to unsafe AI agents	**A potential path to low-risk AI tools**
Open-ended, unguided, recursive improvement	Measured, repeatable, recursive improvement
results in the emergence of a superintelligent system;	yields minimal-content superintelligent learners
the superintelligence gains broad world-knowledge,	that enable systems tutored with specialised knowledge;
develops explicit, long-range goals,	these systems explore solutions to given problems,
develops plans for action with global scope,	perform computations using assigned resources,
employs effective means to implement its plans.	complete assigned tasks by delivering answers.

Note that an essential aspect of part (1) of the low-risk path amounts to standard research practice: storing backups (or checkpoints) of system state during development, and recording the steps that lead to the next interesting result. Together, these practices enable retracing and varying development paths while probing the characteristics of intermediate states.

The following discussion will assume that, along paths toward potentially risky superintelligence, the capacity for recursive improvement precedes strong AI-agent risk, or at a minimum, that this condition can be established by means of controlled redevelopment of recursive improvement capabilities along alternative paths from an

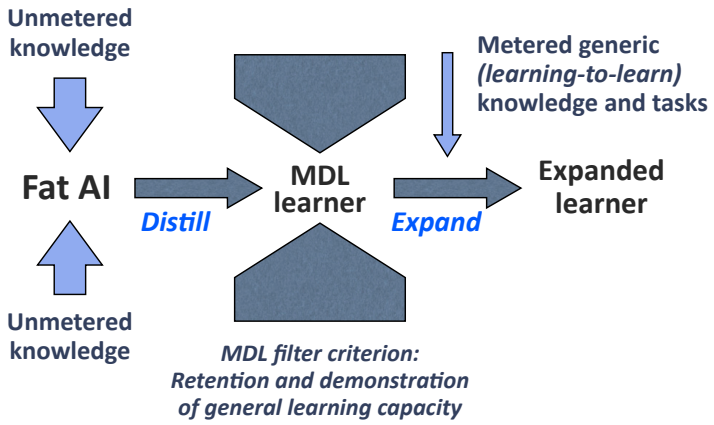

Figure 1. Schematic organization of MDL distillation to produce (and then expand) compact, general-purpose learning systems.

early and non-problematic checkpoint. This condition ensures that control strategies can be applied in a non-adversarial context.

2 Knowledge, learning, and MDL distillation

Along the low-risk path outlined in Table 1, step (2) is pivotal: It calls for the production of a particular kind of superintelligence, a superintelligent learner with minimal information content. How might this be accomplished?

By assumption, the reference problem situation contains AI systems capable of implementing AI systems more intelligent than themselves. A suitably capable base-AI system then can be given as an argument to an AI-improvement operator that applies the base-AI to rewrite a second AI system in order to produce a third, more intelligent AI system:

(1) **Improve**(*base-AI, object-AI,* **metric**(*tasks, smarter*)) ⟶ *smarter-AI,*

where "smarter" is defined in terms of suitably general task-performance metrics.

We can presumably parameterize this operator with any of a range of metrics for improvement, including a metric on the information content of the product:

(2) **Improve**(*base-AI, object-AI,* **metric**(*tasks, smaller*)) ⟶ *smaller-AI.*

Here, improvement entails reducing the size of the product AI conditioned on continued adequate task performance.

The criterial tasks might require that the product AI satisfy a broad range of performance tests *after learning from appropriate curricula.* Given a sufficiently general, superintelligent object AI, a suitably chosen set of criterial tasks can ensure that the product AI system is a general, superintelligent learner.

In the reference problem situation (which assumes an opaque, strongly-improving AI technology), we can apply the improvement operator as follows:

(3) **Improve**(*initial-AI, initial-AI,* **metric**(*tasks, min-MDL*)) ⟶ *MDL-distilled-AI,*

where the resulting "MDL-distilled AI" has two key properties:

1) The task-performance criteria ensure that, like the initial AI, the product is capable of open-ended learning and recursive improvement.

2) The MDL metric ensures that, within resource constraints, the product is the most compact such system that the initial AI could construct.

2.1 Why would an AI system pursue MDL rather than intelligence?

An AI-improving AI system could naturally perform a range of compact-AI implementation tasks, developing MDL-compact versions of systems that can learn to play chess, or learn to beat Watson at playing *Jeopardy!,* and so on. Developing compact versions of systems capable of open-ended learning and recursive improvement is a fundamentally similar kind of implementation task: Optimization of a system for compactness subject to general criteria for learning and performance. Note that tasks of this sort do not entail reflexive, self-modification concerns.

To the extent that concerns might arise regarding problematic strategic behavior in opaque, ill-characterized AI systems, these concerns could potentially be addressed by (for example) restarting an AI-improvement process from a non-problematic checkpoint and interposing MDL distillation steps along the way.

2.2 Omitting language content, omitting domain knowledge

"Knowledge metering"—controlling information inputs—offers a powerful technique for constraining the content of MDL-distilled systems. Consider language:

Infants demonstrate that intelligent systems can achieve general learning capabilities without recourse to an initial endowment of language content (that is, without knowing specific grammar or vocabulary). In particular, general language-learning ability is a consequence of strong priors on abstract language structure in combination with very weak priors on concrete language content.

Distillation of MDL learners would naturally omit vocabulary because vocabulary is MDL-bulky and easily taught or installed. Note that vocabulary cannot be guessed without specific knowledge—would guesses yield a dictionary of Chinese, English, Klingon, or Chicomuceltec? Vocabulary, like other historically-contingent linguistic information (*e.g.,* Figure 2), cannot be inferred from language-independent sources.

Similar remarks apply to the historically-contingent bodies of knowledge that comprise the bulk of the content of most academic fields (*e.g.,* the biosciences), and to

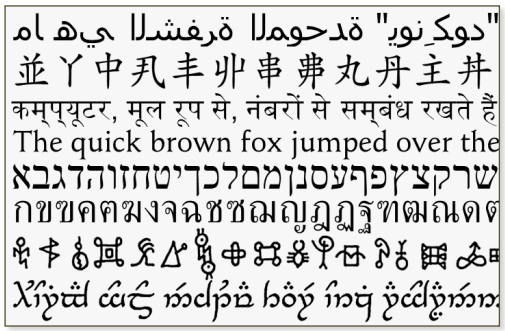

Figure 2. Contingent linguistic information.

knowledge (*e.g.,* of chemistry) that is contingent on physical parameters such as the mass of the electron. The kinds of knowledge that will necessarily (though perhaps implicitly) be retained by an MDL learner presumably fall within the scope of the academic disciplines termed "Formal sciences" in Table 2. The development of professors from infants demonstrates that non-specific priors and general mechanisms provide an adequate basis for open-ended learning.

Given that a body of contingent information has been omitted, suitable constraints on information inputs can preclude its later acquisition. Judging the constraints that follow from a particular knowledge-metering policy will, however, require consideration of not only direct, but also inferred knowledge. Bounds on inference will sometimes be clear, but in considering samples of informal world knowledge, for example, the extent of inferential knowledge may be extraordinarily hard to judge.

Table 2. Academic disciplines relevant to widely differing tasks:

1. Humanities

1.1 Human history
1.2 Linguistics
1.3 Literature
1.4 Arts
1.5 Philosophy
1.6 Religion

2. Social sciences

2.1 Anthropology
2.2 Archaeology
2.3 Area studies
2.4 Cultural...
2.5 Economics
2.6 Gender studies
2.7 Geography
2.8 Political science
2.9 Psychology
2.10 Sociology

3. Natural sciences

3.1 Biology
3.2 Chemistry
3.3 Earth sciences
3.4 Physics
3.5 Space sciences

4. Formal sciences

4.1 Mathematics
4.2 Computer sciences
4.3 Logic
4.4 Statistics
4.5 Systems science

From Wikipedia, "Outline of academic disciplines"

5. Professions

5.1 Agriculture
5.2 Architecture...
5.3 Business
5.4 Divinity
5.5 Education
5.6 Engineering
5.7 Environmental...
5.8 Family...
5.9 Human physical...
5.10 Journalism...
5.11 Law
5.12 Library...
5.13 Medicine
5.14 Military sciences
5.15 Public admin.
5.16 Social work
5.17 Transportation

2.3 Omitting externally-oriented plans

To represent plans requires information, and to the extent that plans are not task-relevant, distillation will tend remove the information that embodies them. In particular, plans that are both specific and oriented toward the external world must contain substantial contingent information that is, as we have seen, unnecessary for general learning capabilities.

One might object that, in an (avoidable) adversarial situation, problematic plans might be embedded in task-relevant structures in ways that, by intention, make them

difficult to identify and remove. A superintelligence-enabled distillation process, however, would presumably be able to employ fresh, compact structures of similar functionality. Needlessly complex structures need not be understood to be discarded.

2.4 Distillation fits current research practice

MDL distillation is intended to separate knowledge from learning capability, and in machine learning today, this separation already holds: Deep learning systems may have surprisingly compact abstract specifications, yet can be trained with gigabytes of data to produce systems with megabytes of opaque, numerical content.

In machine learning, separating knowledge from learning capability is both good science and good engineering:

- Separating knowledge content from learning capability facilitates human understanding of learning processes and their products.
- Training content-free learning systems with known datasets enables reproducibility and benchmarking during development.
- Training content-free learning systems minimizes path-dependent biases and enables diverse applications of particular learning methods.
- MDL principles often improve generalization from training examples to data subsequently used in testing, validation, and applications.

At the threshold of recursive AI improvement, MDL distillation could be applied to separate knowledge from learning capability even if these have become entangled, and can thereby provide a way to retain or recover the scientific, engineering, and safety advantages of current research practice.

3 From MDL distillation to superintelligence-enabled AI tools

Implementing the third step along the proposed low-risk path to AI tools (Table 1) calls for tutoring minimal-content superintelligent learners with generic ("learning-to-learn") and then specialised knowledge to produce specialised, domain-specific AI systems. Figure 3 illustrates the general approach.

Tutoring a distilled, effectively empty MDL learner enables metering (and auditing) the initial knowledge-content of the resulting AI products. This approach mitigates part (2) of the reference problem situation, the potential opacity of the knowledge-content of emerging superintelligent systems. Distillation and knowledge metering can constrain knowledge content regardless of its representation.

Specialised competencies can be narrow, yet powerful; potential examples include superintelligent theorem provers, computer architects, and systems with superintelligent engineering competence in solving the joint structural, mechanical, thermal, and aerodynamic problems of hypersonic aircraft design. Tutoring tightly focused specialists while omitting direct or implied knowledge of language, politics, and geophysics may require attention, but need not always be difficult.

In some domains, tasks will carry potentially significant information about seemingly unrelated aspects of the external world. Information brought by a task stream need not be cumulative, however, because problem-solving systems need not carry forward information from previous instantiations (*e.g.,* checkpoints). A more relaxed policy would enable cumulative learning in the form of canonical representations of task-products such as mathematical theorems, digital circuits, or novel mechanical configurations—in other words, compact representations of task-relevant knowledge.

3.1 Specialisation and composition

Narrow specialists will typically address only parts of problems, sharply limiting their applications in isolation. It is therefore natural to combine narrow domain specialists to build modular systems that, though still specialised, have broader utility.

There are extensive precedents for building broad problem-solving capabilities on specialist foundations, for example:

— Neural systems that combine visual, auditory, and motor cortex.

— Engineering teams composed of diverse human specialists.

— Market economies with extensive division of labor and knowledge.

— Complex software architectures composed of modular components.

As these examples suggest, systems composed of diverse specialists can implement extraordinarily broad capabilities. In the context of AI safety, however, this potential highlights the possibility of composing safe components to build risky systems. Thus, although a systematic exploration of potential superintelligence-based systems can begin by examining means for implementing specialised components, attention then must turn to questions of emergent properties, safety, and risk not only over a range of domains and tasks, but in the context of alternative modular architectures.

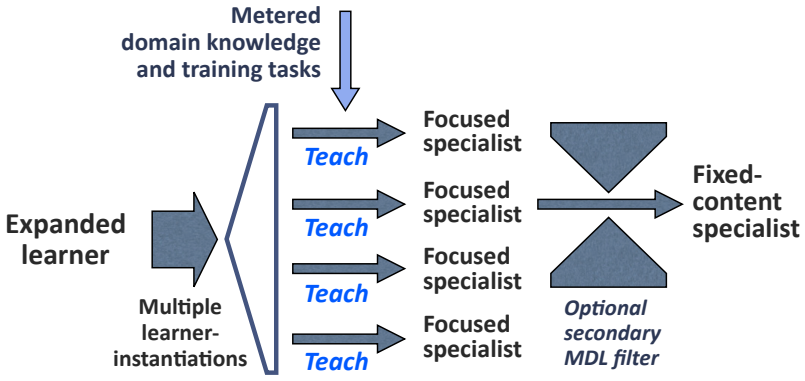

Figure 3. General approach to producing specialist systems from MDL-distilled (then expanded) learning systems (Fig. 2).

3.2 Means and challenges of implementing specialisation

In some areas, knowledge metering can establish clear constraints on potential cross-domain inferential knowledge; in other areas, potential cross-domain inference may be broad and unpredictable. In considering the scope of potential inference, however, it is important to note that the learning capabilities of a specialist system can be curtailed though secondary (post-tutoring) distillation, yielding non-learning systems, and that, as noted-above, cumulative task-related learning can be directly constrained by checkpoint/restart policies.

In addition to constraints on knowledge *per se,* specialist AI can be further shaped and constrained by distillation metrics that optimize resource/performance trade-offs with respect to domain-specific task streams (thereby limiting the scope for other functions), and by fixed interfaces that input task descriptions and output results in domain-specific representations (*e.g.,* mathematical expressions, physical engineering specifications)—in effect, service APIs.

As with inference applied to bodies of knowledge, it will sometimes be difficult to judge the extent of shaping and specialisation that can be induced by task-performance optimization, task-stream control, and domain-specific APIs. These techniques augment a rich set of tools that raise a wide range of questions regarding specialisation, safety, and risk in the context of concrete domains, tasks, and specialist-enabled system architectures.

3.3 Modular specialist architectures

Figure 2 illustrates a general scheme for composing distilled specialists to implement systems with more general capabilities.

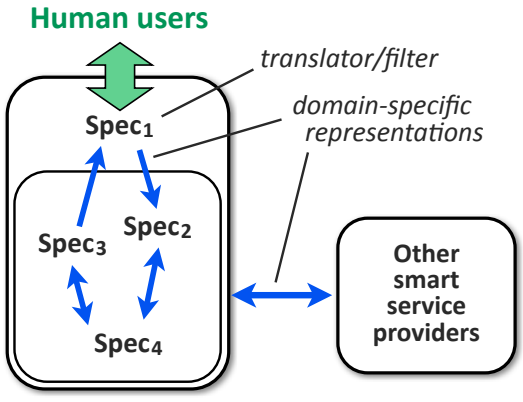

Figure 4. Schematic architecture for a system of linked specialists translated and filtered by a user-interface specialist; see also Fig. 5.

Note that the task of mediating communications between human users and domain specialists might be performed by an interface-communication specialist that enables users to convey and clarify task descriptions through discussion in a domain-specific subset of a natural language (potentially augmented by interactive graphics), while concurrently exchanging task-specific representations with a system composed of domain specialists. Decomposing tasks into narrower subtasks is itself a specialty, as is the translation of results into forms understandable to the human user.

Any or all of these specialists could be made incapable of long-term, cumulative learning by initiating each task with a system in a fixed initial state. To do so would be quite natural: Avoiding task-to-task modifications of system content has the virtue of ensuring consistent behavior, which can be both good engineering practice and an aid to debugging.

The appendix presents a more concrete example of modular specialist composition for the important case of engineering design (Figure 5).

4 Prospects and research directions

Intelligence distillation, knowledge metering, focused specialists, checkpoint/restart and modular composition are general control measures with many potential instantiations and joint applications. These concepts, considered both individually and as a whole, raise questions not only regarding potential scope, implementations, and applications, but also regarding effective methodologies for exploring this range of questions with an eye to potentially critical decisions on paths toward superintelligence.

4.1 Some open questions[2]

How should we interpret "minimum description length"? For practical purposes, a description in terms of Turing machines isn't appropriate; instead, a description might be expressed in a high-level language or executable specification, and might incorporate packaged, opaque algorithms selected from a given library, thus reducing many algorithm descriptions to array indexes. Note that a set of considerations involving resource constraints, curriculum content, and the elastic concept of "learning to learn" are jointly relevant to formulating suitable description-length metrics.

Further, how can we model AI risks and control-measure dependencies? If diverse techniques can be applied to reduce various aspects of strong-agency risk, how can we

---

[2] And a terminological question: "What how should we define 'superintelligence'"? The distilled-learner concept highlights a crucial distinction between *learning* and *competence.* Infants lack adult competence, yet are considered intelligent because of their ability to learn. Accordingly, the term 'superintelligence', as used here, refers both to *superhuman learning capability* and to resulting *superhuman intellectual competencies;* accordingly, use of the term does not imply that any particular system has any particular competencies, human or otherwise.

This contrasts with I. J. Good's 1965 definition of an ultraintelligent machine as "a machine that can far surpass *all the intellectual activities* of any man however clever" (emphasis added). Superhuman learning and superhuman competence must be sharply distinguished.

model risk reduction achieved through multiple techniques? Which control measures can be modeled as probabilistic, independent, and multiplicative? Which are weak if applied separately, yet powerful in combination? Which share common failure modes?

In this framework, what are the thresholds of dangerous agency? What marks the boundary between low-risk AI tools and high-risk AI agents? When might inference from a knowledge base produce unexpected knowledge, and perhaps unexpected capabilities? How broad are the regions that can confidently be regarded as safe?

Prospects for safe applications of superintelligence suggest further open questions:

— How might we exploit a superintelligent theorem-prover?

— What questions could specialised superintelligent systems answer?

— Could superintelligent assistance help us solve AI value-problems?

— Could we structure multilateral games among *untrusted* superintelligent systems to obtain trustworthy solutions to problems of strong AI agency?

Table 3 outlines a sampling of technical topics in need of further exploration. These range from techniques for monitoring capabilities during AI development through specific AI control measures and the scope of their applicability.

Turning to concerns of a different sort, Table 4 outlines a range of considerations related to potential AI development paths, and in particular, key concerns that can be expected to arise in the context of ongoing research and development projects, including the potential costs, uncertainties, constraints, and delays incurred by implementing alternative safeguard policies. The approach to interim AI safety outlined here suggests the possibility of developing concrete and palatable advice that aligns with existing research practice—in particular, methods that separate learning capabilities from learned content—while offering the potential for identifying low-risk paths to a range of rewarding applications of superintelligent AI technologies.

Table 3. A range of technical topics and considerations:

Potential AI-threshold concerns	*Specialist architectures*
• Monitoring emerging capabilities	• Competence factoring
• Applications of checkpoint/restart	• Modular composition patterns
Distillation processes and metrics	*Designing information interfaces*
• Applications of iterative distillation	• Filtering at human interfaces
• Secondary domain-specific distillation	• Monitoring at internal interfaces
Domains and curricula	*Application-specific risks*
• Generic *v.* specialised curricula	• World-interactive robotics
• Teaching *v.* database-loading	• Internet access and interaction
Knowledge partitioning	*Risks of agency*
• Domains and partitions	• Boundaries of risky agency
• Knowledge-scope ambiguities	• Safe composition of risky agents

Table 4. A range of AI-development research considerations:

Current AI research practices	*Expected economic concerns*
• Assessing current practice	• Reducing safeguard uncertainties
• Distillation as good science	• Minimizing safeguard costs
• Assessing current applications	• Minimizing safeguard delays
• Precursors of risky AI agency	• Enabling safe applications

Turning to concerns of a different sort, Table 4 outlines a range of considerations related to potential AI development paths, and in particular, key concerns that can be expected to arise in the context of ongoing research and development projects, including the potential costs, uncertainties, constraints, and delays incurred by implementing alternative safeguard policies. The approach to interim AI safety outlined here suggests the possibility of developing concrete and palatable advice that aligns with existing research practice—in particular, methods that separate learning capabilities from learned content—while offering the potential for identifying low-risk paths to a range of rewarding applications of superintelligent AI technologies.

Turning to AI-risk research, studies of transitional AI risk management could potentially help to bridge a gap, not only in actual risk control techniques (*e.g.,* the lag in preparedness that defines the reference problem situation, Section 1), but also between the risk-oriented and development-oriented AI research communities. These communities have substantial contact today, yet the bridge between them could perhaps be strengthened.

Risk research focused on the unsolved problems presented by superintelligent AI agency is by nature abstract and long-term, and hence has few actionable implications for the concerns of AI developers today. Inquiry into transitional AI safety strategies (Table 5), by contrast, focuses on exploring the territory between today's research objectives and longer-term concerns; it could offer advice relevant to near-term concerns, and could perhaps help us to reframe and reformulate problem situations for research into long-term AI risk control.

Table 5. A range of AI-safety research considerations:

Bridging a gap in AI research agendas	*Addressing long-term objectives*
• Near-term *v.* Long-term concerns	• Enriching the conceptual universe
• Concrete *v.* Abstract problems	• Seeking paths through the transition
• Applications *v.* Risk research	• Seeking enablers for full solutions
Broadening support for risk research	
• Engaging new researchers	
• Addressing a wider range of problems	
• Motivating a wider range of funders	

5 Summary

In the familiar and challenging reference problem situation, AI technology has reached the threshold of rapid, recursive improvement based on opaque, poorly-understood AI systems, while economic and other pressures ensure the application of emerging superintelligence to practical problems before solutions to the problems of strong AI agency are known and implementable.

To address this potential situation, a key aim of transitional AI-risk reduction techniques is to enable applications of superintelligence while minimizing the risks of AI agency. To the extent that transitional AI risk management can delay those risks while providing safe access to powerful intelligent resources, it can contribute to solving the more fundamental problems in several ways: by buying time for further research, by informing research with concrete experience, and, perhaps, by enabling us to use superintelligent problem-solvers to help us solve the problems of superintelligent agency.

To address risks in the reference problem situation, superintelligent AI-improvement capabilities could be applied to the task of producing (distilling) the simplest possible general-purpose learners, defining simplicity by a suitable minimum description length metric. MDL-distilled learners developed by means of appropriate protocols can with high confidence be assumed to lack significant domain knowledge in areas not directly related to successfully performing a set of criterial learning tasks.

Uncertainties regarding the content of what are still (by conservative hypothesis) opaque AI systems can be constrained by training multiple instances of MDL-distilled learners with focused, audited knowledge comprising curricula for distinct specialties. Secondary distillation can further narrow retained knowledge to the essentials required for subsequent domain-specific yet qualitatively superintelligent task performance; as a further knowledge-metering measure, checkpoint/restart protocols can preclude cumulative learning from subsequent task streams.

Finally, the primary limitations of narrow domain specialisation can be addressed by composing narrow capabilities to form more comprehensive systems. Suitable architectures can enable systems to address problems that include communication with human users while restricting the incorporation of general information about the world.

Table 6. A set of composable techniques for transitional AI risk management:

Intelligence distillation	to control initial information content
Knowledge metering	to control information input
Checkpoint/restart	to control information retention
Focused curricula	to train narrow domain specialists
Modular architectures	to compose specialists for practical tasks

In this connection, the following appendix explores a potential architecture for interactive, AI-enabled engineering in more depth.

Table 6 summarizes a set of techniques, that, in creative and careful composition, could provide a powerful approach to shaping the content and functional capabilities of superintelligent AI systems.

In themselves, these techniques cannot ensure safety, because the modular composition of specialist AI systems could be used to implement systems with emergent and effectively unconstrained superintelligent capabilities. Although criteria for reliably safe AI applications are not yet well understood, one can nonetheless anticipate that well-chosen strategies employing these techniques could substantially expand the range of recognizably safe terrain.

Finally, looking beyond incremental extensions of safe AI applications, perhaps the most important motivation for pursuing this line of research is the possibility that strategies for safely applying superintelligent problem-solving capabilities could point the way to strategies for applying superintelligence to solving the fundamental problems presented by superintelligent agency.

Appendix: Safe architectures for superintelligent engineering

Superintelligent AI-based engineering is important not only for its potential applications, but also as an example in which the roles of specialisation, modularity, and task-composition are strong and relatively well understood.

Highly functional AI-enabled engineering systems should:

— Discuss design requirements with users
— Generate candidate designs
— Test candidate designs in simulation
— Evaluate design performance
— Present and explain designs to users
— Iterate design-cycles as necessary
— Remember design discoveries

The architecture outlined here suggests how these capabilities might safely be provided by means of a modular composition of specialists, and it accordingly outlines a task decomposition that would enable user interaction, iterated design and evaluation, and cumulative domain-specific learning (in effect, memoization).

Figure 5 diagrams a proposed coarse-grained task decomposition and associated interfaces.

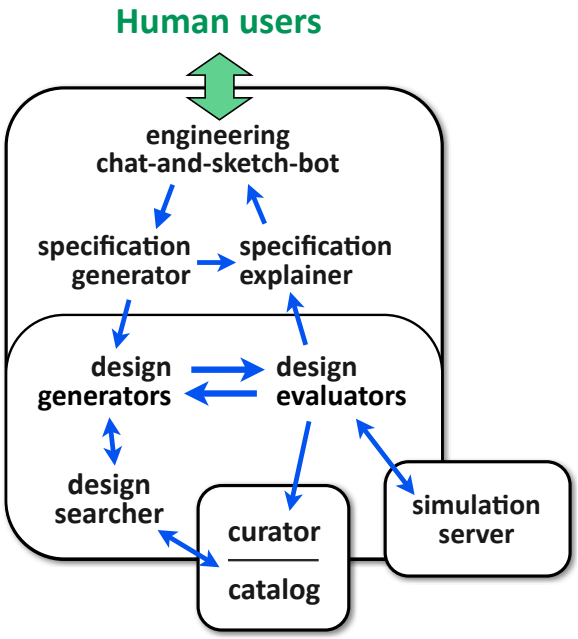

Figure 5. Distilled specialists composed to implement a system with scope for broad engineering competence.

A.1 Human-interface subsystem

In this conception, the human-interface subsystem consists of two layers of specialists: The outward-facing portion is a "chat-and-sketch-bot" that serves as a smart, interactive, human interface with joint competence in domain-relevant language and diagrams. Its knowledge content is specialised with respect to an engineering domain, a user's language, preference settings, and so on. The inward-facing interface of the chat-and-sketch-bot produces annotated diagrams, tables of performance criteria, and the like.

These diagrams, tables, *etc.,* are passed to a "specification-generator" that produces a formal, essentially physical description of the engineering task; an inverse "explanation generator" translates physical descriptions into forms that the outer specialist can present to a user. In iterative task specification, the explanation generator might report requirements that the specification-generator flagged as ambiguous, inconsistent, or cannot be satisfied.

This multi-component human-interface subsystem plays no role in engineering tasks *per se:* The engineering competence of the system depends on the contents of the inner box in Figure 5.

A.2 Specialised engineering subsystems

Problem-solvers for engineering tasks can be decomposed into candidate-design generators and candidate-design evaluators; the latter components test and score designs with respect to physical constraints, criteria, and performance metrics.

Figure 5 diagrams a system at this level of abstraction, including the potential for generative processes to draw on catalogues of previous solutions to design problems, and that evaluation processes can employ external specialists in physical modeling and simulation, and can also, from time to time, transmit designs to a catalogue-curator. The curator stores and indexes designs that meet criteria for novelty and performance. (Note that storing designs in the form of canonical, parameterized, MDL representations can not only reduce their information content, but will typically expand their generality of application and facilitate search.)

Enabling catalog-mediated storage and retrieval of designs can implement an effective and yet narrowly domain-specific form of cumulative learning. In effect, catalogue-mediated storage and retrieval allows systems to learn and share a growing set of (provisional) if-then rules for engineering design; alternatively, storage-and-lookup can be viewed as a form of memoization.

A.3 System architectures

Figure 5 diagrams engineering systems at a high level of abstraction and aggregation; in practice, an engineering system would be implemented as a finer-grained network of subsystems. In engineering, form follows function, both in designed products and in

design processes; it would be natural to template engineering-system processes and architectures on familiar patterns of task specialisation in engineering organizations.

Task organization in engineering (for all but simple or repetitive tasks) involves top-down, hierarchical decomposition of system requirements into subsystem requirements, and (lower levels), decomposition of design tasks into specialties such as optics, structures, electronics, and so on. Functionally, each relationship in this organization entails iterative, bidirectional exchange of domain-specific representations of tasks and candidate solutions, because iterative design generation and evaluation are characteristic of engineering design tasks.

Much more could be said about potential architectures and applications of AI-based engineering systems, but the above description gives a sense of the abstract relationships among task structures, specialisation, and learning.

The nature of specialist roles in engineering may give a more concrete sense of how MDL learners might be used to produce specialists by tutoring learner-instances with focused domain knowledge and tasks, followed by domain-specific secondary distillation.

A.4 Safety considerations and generalizations

Considering the architecture outlined above, there seems good reason to think that the techniques of intelligence distillation, specialisation, and architectural modularity could enable a range of engineering systems, performing at a superintelligent level, to be developed and applied safely, which is to say, employed without incurring a substantial risk of problems involving strong AI agency.

Connoisseurs of subtle AI-risk mechanisms will recognize that systems developed and applied in *formal* accord with the template outlined above could nonetheless present unacceptable intrinsic risks: Modes of means-ends analysis form a continuum, and viewed abstractly, that continuum embraces both circuit design and strategic planning.

By the same token, however, it may be fruitful to explore generalizations of superintelligent engineering systems, pursuing a closer analysis of potential architectures, applications, risks, and risk-countermeasures. Strategies that exploit specialist problem-solving architectures can be expected to generalize across a wide range of AI tasks, and understanding the scope of these strategies could potentially contribute to solving a correspondingly wide range of problems involving the safe application of superintelligent problem-solving capabilities.

Acknowledgements

This work has benefited from my conversations with many members and affiliates of the Future of Humanity Institute, including Nick Bostrom, Stuart Armstrong, Owen Cotton-Barratt, Paul Christiano, Daniel Dewey, Toby Ord, and Anders Sandberg, and in its formative years, has been shaped by conversations with Mark Miller, now at Google Research, and with my then MIT advisor, Marvin Minsky.