Olivia Lee
Stanford Existential Risks Initiative
Winter 2021

Implications of the Comprehensive AI Services Framework on AI Safety Research

Part 1: Introduction

The Comprehensive AI Services (CAIS) framework is a proposal detailed in *Reframing Superintelligence* (Drexler, 2019) that more concretely conceptualizes the progress of AI development towards a superintelligent system. Contrary to the traditional idea of a singleton superintelligent AGI agent, Drexler forecasts the development of a suite of AI services. Each AI service is a system that delivers bounded results using bounded resources in bounded time for some specific task, and each has superintelligent capabilities within its narrow objective or task. He argues that the majority of complex problems humans face can be broken down into tasks, and thus a comprehensive suite of services would theoretically be able to complete a wide range of tasks, creating a generally intelligent system. In light of arguments for potential existential risks posed by AI in the long term (Bostrom, 2014) (Russell, 2015), this paper argues that developing powerful AI systems in line with the CAIS framework is not just likely but should be encouraged, due to the potential for enhanced safety measures to mitigate AI existential risk.

This paper presents the following three hypotheses:

(1) It is plausible that powerful AI systems will be developed in a manner outlined in *Reframing Superintelligence* (Drexler, 2019) in the nearer term, and the CAIS framework is potentially more immediately realizable for future developments in AI.

(2) It is easier to make safer, more interpretable, and more transparent agents using CAIS compared to a traditional singleton AGI. The application of concrete safety tools can contribute to mitigating the existential risk posed by powerful AI systems.

(3) Most previous work in AI safety has, to varying degrees, assumed that powerful AI systems will be developed as singleton AGI systems. In light of (1) and (2), below are some proposals for future research areas that warrant particularly close attention:

(a) safety benchmarking tools: These tools will be applied pre-deployment to ensure safe and comprehensive simulation techniques, demonstrating that the AI system generates the correct outputs for a wide variety of valid inputs, including specific edge cases.

(b) transparency and interpretability tools: These tools will be applied both pre-deployment and post-deployment. Pre-deployment transparency and interpretability tools improve our ability to predict the impact of the service(s) after deployment. Post-deployment tools would speed up the process of safety checking and determining if and when models need to be temporarily stopped and retrained.

(c) monitoring systems: These tools will be applied to monitor the performance of the AI services post-deployment. This is related to the application of the previous category of tools, in order to determine if and when models need to be temporarily stopped and retrained.

(d) hierarchical RL systems: The development of such systems is premised on the development of the previous three tools, especially sound simulation techniques.

This paper will begin with a high level overview of *Reframing Superintelligence* (Drexler, 2019), paying particular attention to the sections detailing the CAIS framework. It then discusses the plausibility of CAIS and engages in a comparative analysis of the risks posed by the CAIS framework and traditional AGI. Finally, it puts forward four opportunities for future research in light of the prior discussion.

Part 2: Hypothesis 1 - Plausibility and Realizability of CAIS

Below is an overview of five key elements of the CAIS model, paying particular attention to the features that are distinct from a traditional singleton AGI agent:

(a) Definition of "service"

A service is an AI system that has a narrow objective or goal, and delivers bounded results for some task using bounded resources in bounded time. Each service has the potential to be superintelligent in its narrow objective, though this does not necessarily make the task simple. For example, superintelligent language translation would count as a service, even though it is a complex task that requires a very detailed understanding of the world and human society. The bounded nature of each service prevents the system from engaging in long-term planning processes, since it is optimizing for a bounded task. The high optimization pressure towards a narrow task allows the system to avoid the problems associated with the standard convergent instrumental subgoals (Bostrom, 2014). Bilevel optimization techniques can be used to bridge the disparate narrow services. The outer optimization task (or upper level task) can be decomposed into a sequence of nested inner optimization task(s) (or lower level task(s)) such that the sequence of optimal solutions of these relaxed problems converges to a potential feasible solution to the upper level optimization problem (Tuy, 1998) (Sinha, 2013). This is distinct from the traditional notion of a singleton AGI agent, which has arbitrary reasoning capabilities that can be applied to a wide variety of problems, similar to general human reasoning.

(b) Distributed system of services

The CAIS framework describes a distributed system of services. This implies that the output(s) of one service in the system is not automatically fed into other services. As a result, the outputs from each service can be individually verified. This is a crucial feature enabling the increased transparency and interpretability of a system constructed using this framework. Unlike an end-to-end singleton AGI system where outputs from several networks are fed from one network to another, with only the final output being interpreted, the interactions between services in the CAIS framework occur through clearly defined communication channels, allowing programmers to more easily backtrace and identify any errors that propagate through the system.

(c) Comprehensive system of services

A vast majority of human activities can be decomposed into smaller tasks. For example, driving a car requires high-speed and high-accuracy visual and auditory processing, decision-making, and fine motor control, each of which can be further decomposed into narrower tasks of higher granularity. Though each individual service has a narrow objective, given a sufficiently broad or comprehensive set of narrow services, one can theoretically complete any task, thus this collection of services in aggregate can be considered generally intelligent. However, in order to reach the state where the system has a sufficiently comprehensive set of services, the system must have a service dedicated to creating new services. Such a service must first recognize that current services are insufficient for a given task, reason about what new service(s) is/are needed to complete this task, and call the relevant existing service(s) or signal to human engineers to develop the new service(s) needed. Therefore, the comprehensive nature of the system is analogous to the idea of generality in artificial general intelligence. Rather than a single end-to-end system that is capable of arbitrary general reasoning, "general" intelligence in the case of CAIS would involve a service that has superintelligent capability in matching the task provided to the specialized service(s) that can perform that task. If there are capabilities required that the system does not currently have, it either initiates or proposes the development of those services to address this need. Ultimately, rather than acting like a single central processing unit that strives to achieve a particular goal, the central processing unit would be more analogous to a search engine, searching through tasks it can perform and calling upon a series of subroutines to achieve the goal.

(d) Separation of R&D and applications

In the CAIS framework, AI R&D and AI applications are separated. The former is driven by humans, though the R&D processes may be (at least partially) automated. This leads to recursive technological improvement, which is distinct from recursive self-improvement in that the improvement arises from improvements in basic AI building blocks which feed back into the R&D services. An analogy to describe this separation is that an autonomous vehicle should be focused on getting the human passenger from origin to destination safely. It should not engage in processes that enable it to become a better autonomous vehicle, for instance by creating new algorithms to improve its performance by speeding up its neural networks. Such processes

should occur under human supervision separately from the period when the system is deployed. This is a crucial feature improving the predictability of systems built using this framework compared to singleton AGI agents that potentially have divergent goals. Such systems render humans powerless if the agent is able to recursively self-improve, increasing the divergence between its own goals and human goals. This separation ensures that human engineers maintain control in reviewing the outputs of individual services in the system, and can thus determine the direction of R&D efforts.

(e) Limited extent of information exchange with the world

The CAIS framework argues for limiting the channels through which the AI systems can influence the world, as an additional layer of control that humans can maintain over AI systems. This is not necessarily limited to physical influence over the world and environment, for example through the embodiment of the AI system. This refers more generally to the extent to which AI systems are able to send and receive information to and from the world and human society. For instance, giving an AI system unfettered access to the Internet gives it tremendous ability to engage in arbitrary interactions with the world and human society, providing it the ability to effect tangible change in the real world through interactions with humans. Limiting the extent to which AI services can exchange information with the outside world will enable humans to preserve greater control over the predictability of the overall system.

Overall, the key argument of the CAIS framework is that general intelligence need not necessarily look like a singleton agent created with the explicit goal of building an agent capability of arbitrary general reasoning. Instead, it may look like a collection of services or product offerings, analogous to the ‘app store model’ in which we have access to a system that is, overall, generally intelligent because of the expansion in breadth and depth of AI services available to us. It predicts that instead of relying on some breakthrough in AI that allows us to achieve general intelligence similar to general human reasoning, narrow AI will continue improving significantly at performing each of its specialized tasks, and the range of tasks that can be achieved by AI services will continue to expand. Once a sufficient number of services have been developed, especially to the level of superintelligent capability, the services that the

overall system can provide will be sufficiently comprehensive so as to resemble general intelligence.

Given the current developments in AI and machine learning, specifically the trends towards superintelligence observed in narrow AI, in comparison to the progress towards an agent that exhibits general reasoning capabilities, it seems plausible that powerful AI systems will be developed in a manner outlined by the CAIS framework. In other words, it is potentially more realistic that we will achieve the disparate capabilities of an AGI agent before we can actually create a singleton AGI agent, and that CAIS systems will be realized earlier than traditional monolithic AGI agents.

Part 3: Hypothesis 2 - Enhanced safety of CAIS systems

Given the key features of systems developed in line with the CAIS framework described above, I propose that CAIS systems are not only more likely to be realized earlier than singleton AGI agents, but are also safer and less likely to lead to catastrophic outcomes.

Firstly, the nature of CAIS systems, which comprise a wide range of narrow services, reduces the possibility of reward hacking leading to unintended outcomes as compared to general optimization problems. The narrow problem definitions of each service makes it easier for engineers to generate trip wires to check individual services and monitor their behavior and outputs. The limited scope of each individual system makes it easier to verify, both formally and experimentally, that the outputs of the behavior are safe and as predicted. Trip wires are plausible vulnerabilities that are deliberately introduced to detect whether an agent attempts to hack its reward function (Amodei et al., 2016). An agent technically has the ability to exploit these vulnerabilities but should not if its value function is correct. If these vulnerabilities are exploited, humans are alerted and the agent can be stopped, thus reducing the risk or at least providing diagnostics about reward hacking behavior.

Secondly, the slower projected takeoff speed of CAIS systems relative to a singleton AGI agent is an additional safety feature. AI takeoff speeds refers to how quickly the production and deployment of AI will be leading up to transformative AI, which is relevant for estimating and mitigating potential risks from advanced AI (Karnofsky, 2016). The distributed nature of the CAIS system and the fact that it does not undergo recursive self-improvement will likely result in a slow-medium speed takeoff situation. The iterative development of individual services, as well as the broadening of scope of the overall system by adding services, will still be largely driven by humans. Furthermore, in the CAIS framework, the system is never fully autonomous, and still requires relatively high human involvement. Unlike a singleton AGI agent, CAIS systems will never become “hands-off” systems in which they are completely autonomous and not reliant on human input whatsoever. Monolithic AGI agents are predicted to have especially fast takeoff speeds because they have little to no couplings to humans. As soon as an AGI agent develops the ability to recursively self-modify or self-improve, it experiences an exponentially fast rate of development. The slower takeoff speed of CAIS systems is crucial to its enhanced

safety. It provides time for any recognition lag that may occur between the onset of unexpected behavior and the identification of this behavior by humans, and the implementation lag in taking steps to correct this behavior or shutting down the service. It also provides time for legislation inertia and global cooperation efforts to understand the development of these systems and come to a mutual agreement on global standards to ensure the safe development of such AI systems, similar to the measures taken to mitigate the existential risk of nuclear warfare. Ultimately, the slow takeoff speed of CAIS systems could potentially help to develop safer AGI and catalyze the enforcement of restrictions on experimentation and development related to AGI.

Overall, rather than the traditional idea of a single, end-to-end, opaque and superintelligent agent that we must try to analyze in advance without really knowing what it will look like or how it will behave, we have a connected network of disparate services. In the event that there is an error in one of the services, or we do not want the system to be able to perform certain tasks for whatever reason, we can simply shut down the individual service(s) and stop the system's access to those services. This eliminates the risk of needing to outsmart or outmaneuver a superintelligent, opaque, end-to-end agent. The nature of the system that responds to complicated tasks by calling upon one or a few of the myriad specialized services that have been developed makes it easier to create safer, interpretable agents using CAIS. While these safety measures do not guarantee that the system will be safe, such measures create multiple checks and safeguards for unsafe behavior, and limits the negative repercussions if unexpected behavior does occur.

Part 4: Hypothesis 3 - Proposals for Future Research

In light of the prior two hypotheses, below are some proposals for future research areas that warrant particularly close attention. A list of AI safety tools exists in (Amodei et al., 2016), however these are four tools that are especially relevant in light of the plausibility of CAIS. One commonality to note that is shared by the four tools is the high level of human involvement - human engineers are highly involved in developing, applying, and interpreting the outputs of these tools, both pre- and post-deployment.

(a) verification tools via simulation

The development of safety benchmarking tools includes safe, accurate, and comprehensive simulation techniques for safe exploration to be applied during model training. Unlike some traditional assumptions about how reinforcement learning agents will be deployed, it is critical that the agent's policy remains static after deployment to ensure the predictability of the model. Simulation techniques in this case can refer to either online virtual simulations or physical simulations in a safe testing environment, or a combination of both. For example, using the analogy of an office cleaning robot (Sutton & Barto, 1998), the safety of the system could first be verified in a virtual simulation (with events controlled by human programmers). The advantage of the virtual simulation is that it is easy to reset the simulation to an initialization state, tweak or modify variables in the simulation, and deliberately put the robot in unusual or "edge case" situations to test its response. It is also easy to investigate situations in which the agent responds in unexpected manners. Once the safety of the system has been ascertained to a certain level, a physical simulation with volunteer test subjects could be implemented, so long as the necessary precautionary measures (specifically the ability to quickly stop the agent) are taken. The advantage of the physical simulation is that it could introduce some "randomness" that is inherent in a real office situation that may not be accounted for in the virtual simulation. For this technique to be successfully implemented, one must ensure that the proxy simulation environment is similar to what the agent will actually encounter in reality. Techniques such as domain randomization (Peng, 2017) can also ensure less distributional shift between simulation and the real world, and prevent the model from overfitting to situations in the training data and focus on the important aspects of the simulation. Such techniques are critical for AI

safety-related research to improve the predictability of individual services in the CAIS framework.

(b) transparency and interpretability tools

A key advantage of CAIS is that interaction between individual systems occurs using clearly defined communication channels. Even though each individual service may be opaque, the system overall is interpretable given that the outputs passed from one narrow subagent to another can be traced. Similar to the concept of decomposition in software engineering, the additional advantage of specifying bounded tasks is that we can backtrace and identify problems in a system by narrowing it down to one or a few subagents, as opposed to writing off an entire AGI agent as incorrectly trained. Since the models comprising the system are static, errors in training of one model will not propagate and compromise successive models which take the outputs of the problematic model as inputs. Capitalizing on the distributed nature of the CAIS system will enable us to circumvent transparency issues with end-to-end singleton AGI agents.

In particular, services dedicated to prediction of outcomes when the AI system has been deployed may be helpful in speeding up the safety verification process. These services should be trained to predict how other component services will behave, giving insight into the potential benefits and risks they bring about. For example, if the system is given a task of mapping a route from origin to destination, a map generator would provide a route and a prediction of how long it will take. A predictive service would check if the predictions provided matched the actual amount of time taken to travel from origin to destination, and can thus be used to detect distributional shifts or divergence in outcome. This would act as a signal to engineers indicating that the model needs to be retrained or updated. These tools will be applied both pre-deployment and post-deployment. Pre-deployment transparency and interpretability tools improve our ability to predict the impact of the service(s) after deployment, thus allowing us to make the necessary modifications to the system before deployment. Post-deployment tools would speed up the process of safety checking and determining if and when models need to be temporarily stopped and retrained. Overall, transparency and interpretability are of paramount importance for intelligent agents to be successfully integrated into human society. A system of narrowly intelligent subagents connected by clearly defined communication channels provides a much

safer alternative to a monolithic AGI agent where the explainability of its actions is significantly more challenging.

(c) monitoring systems

The development of systems to monitor particular metrics post-deployment will be essential in ensuring the system's safety after training. These tools will be applied to monitor the performance of the AI services post-deployment. This is related to the application of the previous category of tools, in order to determine if and when models need to be temporarily stopped and retrained. This is especially important for specific edge cases that may not have been captured during training and ensures the system does not respond in an unexpected manner. If any specific pattern is observed among instances when the system fails to respond, the system can be retrained to account for that specific instance.

(d) hierarchical reinforcement learning (RL) systems

Finally, the achievement of the aforementioned three tools could enable the development of hierarchical reinforcement learning systems. In hierarchical RL systems, a top-level agent takes a small number of abstract actions and completes them by delegating them to sub-agents, which it incentivizes with a synthetic reward signal representing correct completion of the action. These sub-agents themselves delegate to sub-sub-agents, and at the lowest level, agents directly take primitive actions in the environment. The overarching planning agent may be able to learn from very sparse rewards, since it does not need to learn how to implement the details of its policy. On the other hand, the sub-agents will receive a dense reward signal, since they are optimizing synthetic reward signals defined by higher-level agents. Hierarchical RL is still an active research area (Barto & Mahadevan, 2003). With the above developments, it could be possible for an overarching "planning" service in the CAIS framework to distribute roles among smaller systems, which themselves recursively delegate to RL sub-systems at the more stochastic lower levels. This is similar to previously proposed hybrid approaches to hierarchical RL involving RL teleo-operators (RL-TOPs) (Ryan & Reid, 2000). The idea of RL-TOPs is in line with the comprehensive nature of the CAIS framework described in Part 2(c), in that the central processing unit of the system would be more analogous to a distributive system that searches through tasks it can perform and calls upon a series of subroutines to achieve the goal.

Part 5: Conclusion

In this paper, I have put forward three hypotheses related to the realizability and enhanced safety on the CAIS framework in comparison to a traditional singleton AGI unit, as well as four opportunities for future research in light of the prior discussion. Ultimately, a connected system of AI subagents reduces the risks associated with low transparency, interpretability, and predictability of end-to-end singleton agents, while also yielding tangible benefits through the division and specialization of narrow functions in the suite of AI services. Overall, this contributes to reducing the likelihood of AI agents becoming a significant destabilizing force in society, because in this framework, humans are still actively involved in the construction, operation, and monitoring of AI systems. The process of testing, training, and deploying AI services is still ultimately in the control of human programmers. This system reflects the interdependence between humans and machines that we should strive towards to mitigate AI existential risk, as opposed to a system of complete autonomy and potential replacement that arises from the development of monolithic AGI agents.

Bibliography

Amodei, D., et al. (2016). *Concrete Problems in AI Safety*. http://arxiv.org/abs/1606.06565

Barto, A., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Special Issue on Reinforcement Learning, Discrete Event Systems Journal*, 13, 41–77.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

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

Karnofsky, H. (2016). *Some Background on Our Views Regarding Advanced Artificial Intelligence*. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence

Peng, X. B., Andrychowicz, M., Zaremba, W., Abbeel, P. (2017). *Sim-to-Real Transfer of Robotic Control with Dynamics Randomization*. https://arxiv.org/pdf/1710.06537.pdf

Ryan, M. R. K., & Reid, M. D. (2000). Using ILP to improve planning in hierarchical reinforcement learning. *In Proceedings of the tenth international conference on inductive logic programming*, ILP 2000, London. London: Springer.

Russell, S., Dewey, D., Tegmark, M. (2015). *Research Priorities for Robust and Beneficial Artificial Intelligence*, AI Magazine, Vol. 36, No. 4.

Sinha, A., Malo, P., Deb, K. (2013). *Efficient Evolutionary Algorithm for Single-Objective Bilevel Optimization*. http://arxiv.org/abs/1303.3901

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tuy H. (1998) Outer and Inner Approximation. *Convex Analysis and Global Optimization. Nonconvex Optimization and Its Applications*, vol 22. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-2809-5_6