AI Vision and Future - Appendix

AGI Foundational Technologies Research and Product Sources and Source Code Repositories

Gene Stevens
Fall 2025

World Models & Simulation

Overview

World Models & Simulation form the backbone of AGI's ability to understand, predict, and interact with reality. For AGI to move beyond narrow tasks, it must internalize causal, physical, and social dynamics -- much like humans form mental models of how the world works. By simulating environments, AGI can safely learn from trial and error, extrapolate outcomes, and prepare strategies for real-world action without excessive risk. This domain is critical because it enables generalization, causal reasoning, and transfer learning -- key capabilities that distinguish AGI from specialized AI.

---

1) Self-Supervised World Models

Learn representations of the environment without labeled data, discovering structure and dynamics from raw input. By compressing sensory streams into latent spaces, agents can "dream" possible futures and plan accordingly.

- **Key papers**
  - DreamerV3: Mastering Diverse Domains through World Models -- Hafner et al., 2023.
  - MuZero: Planning with a Learned Model -- Schrittwieser et al., *Nature*, 2020.
  - World Models / Presentation Ha & Schmidhuber, 2018.
  - A Generalist Agent (Gato) -- Reed et al., 2022.
- **Code repositories**
  - DreamerV3 GitHub
  - MuZero General GitHub
  - World Models GitHub
  - Unofficial Gato GitHub

---

2) Causal Inference & Causal Graphs

Helps AGI move beyond correlation to understand cause-effect relationships, enabling robust predictions and interventions.

- **Key papers**
  - Causality: Models, Reasoning, and Inference -- Judea Pearl, 2000/2009.
  - What Counterfactuals Can Be Tested -- Shpitser & Pearl, 2008.
  - DoWhy: End-to-End Causal Inference -- Microsoft Research, 2019.
- **Code repositories**
  - DoWhy GitHub
  - CausalNex GitHub
  - Pyro GitHub

---

3) Neural-Symbolic World Models

Combine neural learning with symbolic reasoning for generalization and interpretability.

- **Key papers**
  - DeepProbLog -- Manhaeve et al., 2018.
  - Logic Tensor Networks for Semantic Image Interpretation -- Donadello et al., 2017.
  - Neuro-Symbolic Concept Learner (NSCL) -- Mao et al., 2019.
- **Code repositories**
  - DeepProbLog GitHub
  - Logic Tensor Networks GitHub
  - NSCL GitHub

---

4) High-Fidelity Simulation Environments

Simulation provides safe, scalable, and diverse training/testing grounds for AGI.

- **Key papers**
  - Unity: A General Platform for Intelligent Agents -- Juliani et al., 2018.
  - Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning -- Makoviychuk et al., 2021.
  - CARLA: An Open Urban Driving Simulator -- Dosovitskiy et al., 2017.
- **Code repositories**
  - Unity ML-Agents GitHub
  - NVIDIA Isaac Gym GitHub
  - CARLA GitHub

Long-Term & Episodic Memory

Overview

AGI needs memory systems that last beyond a single context window and that can store, organize, and retrieve experience across time. Long-term semantic memory provides factual/relational knowledge; episodic memory preserves temporally ordered experiences; and working memory bridges them during active reasoning. Together they support continuity of identity, skill accumulation, tool use, and safe adaptation (avoiding “catastrophic forgetting”)--all prerequisites for robust, general intelligence.

Foundational Technologies: Example Tech & Research

1) Vector Databases & Embeddings

Purpose-built stores for dense vectors (text/code/image/audio) enable semantic retrieval, memory augmentation, and retrieval-augmented generation (RAG) at scale.

- **Key papers / docs**
  - FAISS: A library for efficient similarity search (many papers) – Johnson et al. (Meta).
  - Weaviate Documentation – open-source vector DB with semantic search.
  - Milvus Documentation – cloud-native vector DB with RAG examples.
  - Pinecone Documentation – managed vector DB with RAG use cases.
- **Code repositories**
  - FAISS GitHub
  - Weaviate GitHub
  - Milvus GitHub
  - Pinecone Examples GitHub

---

2) Neural Memory Architectures

Neural networks augmented with differentiable external memory (read/write) or specialized controllers to store and recall information across long horizons.

- **Key papers**

    - Differentiable Neural Computer (DNC) -- Graves et al., *Nature* (2016).
    - Neural Turing Machines (NTM) -- Graves et al., arXiv (2014).
- **Code repositories**
    - DNC GitHub (DeepMind Sonnet/TensorFlow)
    - Neural Turing Machines PyTorch Implementation

---

3) Hierarchical Memory Systems

Architectures that extend context and structure memory into short-term caches and long-term stores. They often pair transformers with retrieval or long-range operators.

- **Key papers**
    - Transformer-XL -- Dai et al., “Attentive Language Models Beyond a Fixed-Length Context” (2019).
    - RETRO -- Borgeaud et al., “Improving language models by retrieving from trillions of tokens” (2022).
    - Hyena Hierarchy -- Poli et al., “Towards Larger Convolutional Language Models” (2023).
    - MEMIT: Mass-Editing Memory in a Transformer -- Meng et al., ICLR (2023).
- **Code repositories**
    - Transformer-XL GitHub
    - Hyena GitHub (HazyResearch)
    - HyenaDNA GitHub
    - MEMIT GitHub

---

4) Continual & Lifelong Learning

Methods to learn new tasks without overwriting old knowledge: regularization, replay buffers, and meta-learning approaches mitigate catastrophic forgetting.

- **Key papers**
    - Elastic Weight Consolidation (EWC) -- Kirkpatrick et al., *PNAS* (2017).
    - Meta-Experience Replay (MER) -- Riemer et al., NeurIPS (2019).
    - Gradient Episodic Memory (GEM) -- Lopez-Paz & Ranzato, NeurIPS (2017).
- **Code repositories**
    - GEM GitHub (FAIR official)
    - EWC PyTorch Implementation
    - Continual Learning Research Collection

Reasoning & Planning

Overview

Reasoning and planning enable AGI to solve novel, multi-step problems and adapt flexibly across domains. Unlike pattern recognition alone, reasoning involves symbolic logic, causal inference, abstraction, and hypothesis testing. Planning further integrates these capabilities to sequence actions toward goals, balancing immediate outcomes with long-term objectives. These foundations ensure that AGI can operate autonomously, explain its decisions, and generalize strategies across diverse tasks -- from mathematics and coding to robotics and scientific discovery.

Foundational Technologies: Example Tech & Research

1) Neuro-Symbolic AI

Merges the strengths of neural networks (learning from data) with symbolic systems (logical precision and interpretability). This approach enables AGI to both recognize complex patterns and reason over structured knowledge.

- **Key papers**
  - DeepProbLog: Neural Probabilistic Logic Programming -- Manhaeve et al., 2018.
  - Logical Neural Networks (LNNs) -- Riegel et al., 2020.
  - IBM Neuro-Symbolic AI Overview -- d'Avila Garcez & Lamb.
- **Code repositories**
  - DeepProbLog GitHub
  - LNN GitHub (IBM Research)

---

2) Automated Theorem Proving & Formal Reasoning

Focuses on enabling machines to **prove logic and math statements** with rigor. These systems formalize reasoning, support verifiable proofs, and underpin safety-critical AGI applications.

- **Key papers and Projects**
  - Lean Proof Assistant -- de Moura et al., 2015–present.
  - Rocq Theorem Prover -- Huet et al., foundational since the 1980s.
  - HOL Light Theorem Prover -- Harrison, lightweight higher-order logic.
  - Generative Language Modeling for Automated Theorem Proving -- Polu & Sutskever, 2020.
- **Code repositories**

    - Lean GitHub
    - Rocq GitHub
    - HOL Light Source
    - Lean GPT-f (previous OpenAI experiments repo has been deleted)

---

3) Program Synthesis

Enables AGI to **generate executable code** from natural language or formal specifications. It demonstrates reasoning over abstract syntax trees, constraints, and programming languages -- a key step toward AI as a problem-solving partner.

- **Key papers**
    - Competition-Level Code Generation with AlphaCode -- Li et al., DeepMind (2022).
    - Evaluating Large Language Models Trained on Code (Codex) -- Chen et al., OpenAI (2021).
    - DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning -- Ellis et al., MIT/DeepMind (2020).
- **Code repositories**
    - AlphaCode CodeContests
    - OpenAI Codex Playground
    - DreamCoder GitHub

---

4) Hierarchical Reinforcement Learning (HRL)

Extends reinforcement learning by **decomposing tasks into sub-goals** and organizing them hierarchically. This makes complex, long-horizon problems tractable for AGI.

- **Key papers**
    - Options Framework -- Sutton et al., 1999.
    - FeUdal Networks (FuN) for Hierarchical Reinforcement Learning -- Vezhnevets et al., 2017.
    - HIRO: Hierarchical Reinforcement Learning with Off-Policy Correction -- Nachum et al., 2018.
- **Code repositories**
    - FeUdal Networks PyTorch Implementation
    - HIRO OpenAI Baselines Implementation

---

5) Chain of Thought & Tool Use

Combines **step-by-step reasoning** with the use of external tools (calculators, search engines, APIs). These methods extend AGI's reasoning capability beyond static models into interactive, adaptive problem-solving.

- **Key papers**
  - Chain-of-Thought Prompting Elicits Reasoning in Large Language Models -- Wei et al., 2022.
  - ReAct: Synergizing Reasoning and Acting in Language Models -- Yao et al., 2022.
  - Graph-of-Thoughts: Solving Elaborate Problems with Large Language Models -- Besta et al., 2023.
- **Code repositories / frameworks**
  - LangChain GitHub
  - AutoGPT GitHub
  - Graph-of-Thoughts GitHub

Embodiment & Interaction

Overview

Embodiment grounds AGI in the physical and social world, enabling learning through **action, feedback, and interaction**. By equipping AGI with bodies (robots, avatars, or digital agents) and sensory systems (vision, touch, proprioception), it gains the ability to experiment, adapt, and form richer world models. Interaction -- whether through speech, gestures, or immersive AR/VR -- ensures that AGI can operate in human environments, collaborate with people, and refine its skills in context. This is critical for bridging abstract intelligence with real-world usability.

---

1) Humanoid Robotics

Robots with human-like form factors that combine perception, locomotion, and manipulation. These systems provide testbeds for AGI's embodied cognition. This area is very light on direct published literature and source code repositories.

- **Key papers / docs / products**
  - Tesla Optimus Overview & Tesla AI Day presentations.
  - Agility Robotics Digit -- Commercial humanoid robot for logistics and services.
  - Sanctuary AI Phoenix -- General-purpose humanoid robots for real-world tasks.
- **Code / repos**
  - Agility Robotics Developer Resources
  - Tesla Optimus Demos (Community Repos) (currently empty)

---

2) Tactile & Haptic AI Systems

Enable robots and agents to sense and manipulate the world through touch, force feedback, and fine-grained physical interaction.

- **Key papers / docs**
  - Shadow Dexterous Hand -- Highly articulated robotic hand used in manipulation research.
  - MIT GelSight -- Vision-based tactile sensing technology.
  - DARPA HAPTIX Program -- Haptic prosthetics and AI touch feedback.
- **Code / repos**
  - GelSight GitHub
  - Shadow Robot Interface packages
  - HAPTIX Research Demos (Community) (now an empty set of repos)

---

3) Digital Twins

Virtual replicas of physical systems that allow real-time monitoring, simulation, and optimization. Critical for AGI training, testing, and interaction at scale.

- **Key papers / docs**
  - Siemens Digital Twin Overview
  - NVIDIA Omniverse -- High-fidelity simulation and collaboration platform.
  - Microsoft Azure Digital Twins
- **Code / repos**
  - NVIDIA Omniverse GitHub Samples
  - Azure Digital Twins GitHub
  - Siemens Industrial Edge GitHub
  - Siemens Digital Twins WASM

---

4) Immersive Interaction

AI systems integrated into AR/VR, enabling natural, multimodal user engagement and collaborative environments.

- **Key papers / docs / products**
  - Meta Horizon Workrooms -- Virtual collaboration environment.
  - Apple Vision Pro -- Spatial computing platform for immersive interaction.

    - EmBARDiment: an Embodied AI Agent for Productivity in XR -- Research survey on AI in AR/VR.
- **Code / repos**
    - Meta OpenXR SDK / Samples
    - Apple Vision Pro Developer Kit
    - OpenXR GitHub

Autonomy & Self-Improvement

Overview

Autonomy and self-improvement enable AGI to go beyond passive responses by **setting goals, experimenting, and refining itself**. This foundation involves agentic frameworks, adaptive learning strategies, and self-directed optimization. Through meta-learning, self-play, curriculum design, and neural architecture search, AGI develops the ability to improve continuously, adapt to novel tasks, and evolve new capabilities. These mechanisms form the “flywheel” of iterative growth -- essential for scalable, general-purpose intelligence.

---

1) Agentic AI Frameworks

Platforms and libraries that provide the scaffolding for **autonomous task execution**, chaining reasoning with action in real or virtual environments.

- **Key papers / docs**
    - LangChain Documentation -- Framework for composable LLM applications.
    - CrewAI -- Multi-agent framework for AI collaboration.
    - AutoGPT -- Experimental autonomous agent using GPT-4.
    - OpenAI Operator -- General-purpose agent model (research blog).
- **Code repositories**
    - LangChain GitHub
    - CrewAI GitHub
    - AutoGPT GitHub

---

2) Meta-Learning

“Learning to learn” methods, where models rapidly adapt to new tasks with minimal data by leveraging prior experience.

- **Key papers**
  - Model-Agnostic Meta-Learning (MAML) -- Finn et al., 2017.
  - Reptile: A Scalable Meta-Learning Algorithm -- Nichol et al., 2018.
  - Meta-SGD -- Li et al., 2017.
- **Code repositories**
  - MAML GitHub (Chelsea Finn)
  - Reptile GitHub
  - Meta-SGD PyTorch Implementation

---

3) Self-Play & Evolutionary Methods

Agents improve by competing against themselves or evolving through population-based strategies. This creates scalable pathways to mastery.

- **Key papers**
  - AlphaGo: Mastering the game of Go without human knowledge -- Silver et al., *Nature*, 2017.
  - OpenAI Five: Dota 2 with Large Scale Deep Reinforcement Learning -- Berner et al., 2019.
  - Evolution Strategies as a Scalable Alternative to Reinforcement Learning -- Salimans et al., 2017.
- **Code repositories**
  - AlphaZero General GitHub
  - OpenAI Five Dota2 Blog & Resources
  - CMA-ES GitHub (Evolution Strategies)

---

4) Auto-Curricula & Learning

Progressively structured training schedules that evolve automatically, allowing agents to tackle increasingly difficult tasks.

- **Key papers**
  - POET: Paired Open-Ended Trailblazer -- Wang et al., OpenAI, 2019.
  - Curriculum Reinforcement Learning -- Bengio et al., DeepMind, 2017.
- **Code repositories**
  - POET GitHub (Uber AI Labs)
  - Facebook ReAgent - RL

---

5) Automated Model Architecture Search

Algorithms that design optimal neural network structures automatically, enabling self-directed architectural innovation.

- **Key papers**
  - Neural Architecture Search with Reinforcement Learning -- Zoph & Le, 2017.
  - Auto-Keras: An Efficient Neural Architecture Search System -- Jin et al., 2019.
  - Google AutoML: Practical Applications -- Google Research. (they pulled the paper since it's now a product; this blog is still out there)
- **Code repositories**
  - AutoKeras GitHub
  - Google AutoML Papers & Tools
  - NAS Benchmarks (NAS-Bench-101)

Safety & Alignment Frameworks

Overview

Safety and alignment are critical for ensuring that AGI systems act in ways consistent with human values, intentions, and social norms. These frameworks focus on aligning AI behavior through feedback, interpretability, preference modeling, adversarial testing, and scalable oversight. The goal is to minimize risks, build trust, and provide robust governance for increasingly capable AI systems. Without strong alignment mechanisms, AGI could diverge from human objectives, making this foundation essential for responsible development.

---

1) RLHF and Extensions

Reinforcement Learning from Human Feedback (RLHF) aligns models by optimizing against human-provided preferences. Variants include **Constitutional AI** and **RLAIF** (reinforcement learning from AI feedback).

- **Key papers**
  - InstructGPT: Training language models to follow instructions with human feedback -- Ouyang et al., OpenAI, 2022.
  - Constitutional AI: Harmlessness from AI Feedback -- Bai et al., Anthropic, 2022.
  - RLAIF: Reinforcement Learning from AI Feedback -- Lee et al., Anthropic/OpenAI, 2023.
- **Code repositories**
  - TRL (Transformers RLHF)
  - OpenAI RLHF Examples

---

2) Mechanistic Interpretability

Understanding model internals by analyzing neurons, circuits, and activations. Provides transparency into why models behave as they do.

- **Key papers**
  - Transformer Circuits -- Olsson et al., OpenAI/Anthropic, 2022.
  - Activation Patching for Causal Interpretability -- Meng et al., 2022.
  - Neuronpedia -- Open database of neuron-level analyses.
- **Code repositories**
  - TransformerLens GitHub
  - OpenAI Interpretability Research GitHub
  - Neuronpedia GitHub

---

3) AI Red-Team & Adversarial Testing

Stress-testing AI against failure modes and adversarial inputs to expose weaknesses before deployment.

- **Key papers / docs**
  - ARC Evals -- Alignment Research Center's evaluation protocols.
  - Anthropic Red Teaming Research -- Systematic adversarial testing.
  - CleverHans: Adversarial Examples Library -- Papernot et al., 2016.
- **Code repositories**
  - CleverHans GitHub
  - Adversarial Robustness Toolbox (ART)
  - ARC Evals Resources

---

4) Value Alignment & Preference Learning

Teaching AGI human values, goals, and social preferences to guide decision-making.

- **Key papers**
  - Apprenticeship Learning via Inverse Reinforcement Learning -- Abbeel & Ng, 2004.
  - Cooperative Inverse Reinforcement Learning (CIRL) -- Hadfield-Menell et al., 2016.
  - TAMER Framework: Interactive Shaping of Agent Behavior -- Knox & Stone, 2008.

- **Code repositories**
  - CIRL GitHub Implementations
  - Inverse RL PyTorch
  - TAMER Framework GitHub

---

5) Scalable Oversight

Using AI systems to evaluate and supervise other AI, enabling oversight to scale alongside capability growth.

- **Key papers**
  - AI Safety via Debate -- Irving et al., OpenAI, 2018.
  - Recursive Reward Modeling -- Leike et al., OpenAI, 2018.
  - AI-Assisted Evaluation for Language Models -- Bowman et al., Anthropic, 2023.
- **Code repositories**
  - AI Debate Model GitHub (community)
  - OpenAI Recursive Reward Modeling
  - Anthropic Evals Resources