Definitions for Conditions

Capability and generality: (Dimension)

1) **Low**: AI Systems remain approximately as capable and general as current systems and progress only marginally with respect to power or general-purpose capabilities. Decreased investment and another AI winter are possible
2) **Moderate**: AI systems become increasingly powerful and generalizable across multiple cognitive tasks in a range of fields. Society struggles to keep pace with changes.
3) **High Narrow**: Systems develop to human and superhuman capable across most tasks, but domain-specific only, with a mix of agents and services (see: Comprehensive AI services model - **CAIS**)
4) **High general:** AI systems progress to human-level AGI. The system is as capable and general as humans in all domains. With computational advantages, the AGI is capable of recursive self-improvement and progress to **ASI**. (Standard hard takeoff "foom" scenario).

Takeoff speed: :(Dimension)

1) **Slow**: (decades or longer) AI systems develop incrementally, with incremental progress. The possibility of an AI winter is high. Powerful capabilities are theoretically possible, but they'd develop over much longer time horizons.
2) **Moderate (uncontrolled/continuous)**: (many months to years, less than a decade) Systems develop rapidly but with no sharp discontinuity. The changes spread faster than anticipated with surprising rapid capability jumps that are extremely difficult for society to keep up with. (see: Christiano's what failure looks like: https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like both versions are included in this condition.
3) **Moderate (controlled/competitive)**: (many months to years, less than a decade) Systems develop rapidly (no sharp discontinuity). Radical changes are anticipated and actively pursued, for competitive advantage, including the lead-up or response to conflict. Unexpected capability jumps but control efforts are planned. This scenario is related to highly competitive race dynamics and could have geopolitical dimensions.

Distribution: (Dimension)

1) **Wide distribution**: AI systems are widely available through open-source networks when HLMI is developed. Resource requirements are low, bringing inordinate power to citizens.
2) **Moderate Distribution**: AI discoveries are made across leading companies only, with technological parity and resources, in several countries. Multipolar scenario.
3) **Concentrated**: The system is discovered by and confined to one lab or government program. This includes scenarios where the discovery is part of a corporation's special program (e.g., Google X), a surprise discovery, or an accident.

Timeframe: (Dimension)

1) **Less than 20 years:** High-level machine intelligence or a close approximation is developed before 2040. The system is capable of completing most cognitive tasks of a human being. This includes the possibility of AGI or ASI but does not depend on that exact instantiation.
2) **20 to 50 years:** High-level machine intelligence or a close approximation is developed before sometime between 2035 and 2070 The system is capable of completing most cognitive tasks of a human being. This includes the possibility of AGI or ASI but does not depend on that exact instantiation.
3) **Greater than 50 years:** High-level machine intelligence or a close approximation is developed before 2040. The system is capable of completing most cognitive tasks of a human being. This includes the possibility of AGI or ASI but does not depend on that exact instantiation.

Accelerants: (Dimension)

1) **Compute overhang or bottleneck:** A new algorithm, overlooked insight, or paradigm exploits existing compute far more efficiently than previously, allowing rapid gains in capability or generality.

2) **Innovation:** A new insight, machine learning paradigm, or completely new architecture accelerates capabilities, from 0 to 100, allowing faster and more general capabilities. Examples could include insight from neuroscience, a new mode of learning (e.g., common sense), or quantum materials or computation.
3) **Embodiment/Data source:** Simulated or actual embodiment, a new type or quality of data for ML training provides radical capability gains.

AI Paradigm: (Dimension)

1) **Current paradigm:** The current machine learning paradigms can scale up radically to advanced capabilities and broad generality, up to and including AGI ("prosaic AGI" - https://ai-alignment.com/prosaic-ai-control-b959644d79c2)
2) **New Paradigm:** HLMI requires an entirely new AI paradigm. New modes of learning such as system two reasoning, a fundamental insight on intelligence, or new architectures, are required to reach high-level general decision making.
3) **Current paradigm plus:** HLMI systems are attainable using current machine learning paradigms but require additional methods. Current learning methods are on the right track but require additional learning techniques, such as a hybrid approach, common sense reasoning, genetic algorithms plus self-supervised learning plus common sense.

Race dynamics: (Dimension)

1) **Cooperation:** AI technologies are recognized as a global public good and cooperation increases between companies and national governments. Race to the top scenario.
2) **Isolation:** Global governments take a protectionist turn and cooperation decreases. AI is developed in isolation. Markets attempt to maintain the status quo and companies compete regionally or within national borders, causing wide disparities in technical standards and regulations.
3) **Monopolization:** Technology companies increase acquisitions of smaller companies and talent to control AI resources. Corporations increasingly control the direction of research, influence over governments, and distribution of power. In the extreme end, companies become semi-sovereign entities beyond the reach of government and international institutions.
4) **AI Arms Race:** AI is named a strategic national asset and countries race for global dominance. As high-level capabilities become more likely, governments begin to control research and access and use top companies as an arm of military power. AI is militarized and conflict is more likely.

Highest Threat Risk Class: (Dimension)

1) **Misuse:** Alignment is under control and Cyber-attacks and disinformation campaigns increase in frequency and disruptive potential. Persistent surveillance becomes more likely by governments and criminals.
2) **Accidents or failures:** AI systems are given more control over decision processes making failure modes more consequential and goal alignment remains the key danger. With systems in control of increasingly sensitive infrastructure, a failure could result in cascades of follow-on failures.
3) **Structural:** Increased decision autonomy of AI systems brings subtle changes to the functioning of society and uncertainty of conflict. Overlap between nations' offense/defense balance makes it more likely for military escalation. Values decline as AI takes control of all decision processes.

AI Safety: (relationship to capability) (Dimension)

1) **Current techniques scale to HLMI**: Current AI safety techniques can scale to high-level systems. The current techniques being designed for the dominant paradigm are broadly transferable to HLMI.
2) **New techniques**: New AI safety techniques must be developed from first principles to be effective against high-powered more general systems.
3) **Custom Techniques**: Each unique instantiation of an advanced AI system requires a specialized safety technique to be developed, making alignment a far more complex problem.

AI Safety Risk: (Dimension)

1) **Goal Alignment**: Goal alignment remains the primary intractable problem that we are unable to solve. Progress in alignment has had success, but system changes require entirely new solutions. The most dangerous risk from HLMI remains misaligned systems.
2) **Power-seeking and deception:** The most prevalent and dangerous concern turns out to be the acquisition of resources by AI systems. Even with improvements to goal alignment, instrumental objectives, and deception to prevent changes, it is difficult to detect and vary across all systems. The potential to lose control is high.
3) **Mesa Optimization**: Goal alignment has had significant success, but inner aligned agent models remain a problem and are extremely difficult to identify. Subtle and impossible to detect misalignment issues and failures remain prevalent and are the most dangerous concern.

Developer: (Dimension)

1) **Coalition of states (e.g., EU, NATO):** A coalition of nation-states, international organizations, or military alliances develop the first radically capable advanced AI systems.
2) **Country:** An individual government discovers or develops radically transformative AI systems. This could be through a national government program, the military, or by nationalizing one or several corporations.
3) **Corporation/Academia:** A private-sector corporation (e.g., Tencent, Google), non-profit, or academic research institution develops the first advanced AI instantiation.
4) **Individual:** A private developer discovers an advanced AI capability. This is more likely in circumstances where AI research and development remains open-source and resource requirements are low (e.g., a new AI paradigm).

Governance: (Dimension)

1) **Weak (decrease in governance):** Preparation stays the same as today (reactive) or decreases in cooperation, collective action, and agreements due to isolationism or conflict and weakening of norms and institutions, possibly due to race dynamics.
2) **Moderate:** A strengthening of international norms and consolidation of institutions. International norms on the proper use of AI systems are well established and an agreed-upon framework of safety standards is established.
3) **Strong:** International safety regimes established (e.g., IAEA), multilateral agreements, and verification measures (e.g., IAEA nuclear inspections) enacted for states unwilling to sign on to AI safety agreements. An international body on AI safety is established that coordinates efforts.

Corporate Governance: (Dimension)

1) **Decrease**: An increase in economic competition brings decreased cooperation across leading AI companies, impacting safety coordination. Isolation could worsen this.
2) **Moderate improvement**: AI companies and research institutions increase coordination on AI development and technical safety practices, with intercompany working groups on technical safety standards and control measures.
3) **Strengthen**: AI companies and research institutions agree on third-party safety standards and a common framework for technical safety control measures.

Developer Location: (Dimension)

1) **USA-Western European:** Major companies in the US or headquartered in the US or the EU develop the first HLMI instantiation. This region additionally includes close allies often considered "western" such as Australia and Japan.
2) **Asia-Pacific:** Greater Asia – South, Southeast, Southwest, and East – develop the first HLMI instantiation. This includes the pacific islands, Eurasia, Russia, and the Middle East.
3) **Africa or Latin America/Caribbean:** The global south, besides Asia. This includes Central, South America, the Caribbean, and continental Africa.

Superintelligence Scenarios: (Dimension)

1) **The internet as emergent intelligence:** Unable to recognize the qualitatively different forms of intelligence, the internet has been developing intelligence as a large complex system. The collective system sparks the emergence of a single intelligence.

2) **Cognitive Internet-of-Things**: As AI is networked throughout all sensors and systems, with the breakdown of the cyber and physical environment, machine agents proliferate across global networks as a sensor web of millions of independent agents, with independent alignment risks.
3) **Narrow AI systems convergence:** As tool AI continues to spread and increase in power (CAIS model), like strands of DNA, these individual agents combine and emerge as one superintelligence (much like a swarm of honeybees becoming the hive).