TOWARDS COMMON SENSE AI: A PHASED APPROACH

Hugo Latapie
hugo@taijituai.com

ABSTRACT

Recent announcements from Apple, Amazon, and OpenAI underscore a recognition that simply scaling data, model sizes, and compute does not automatically deliver reliable and beneficial AI solutions for high-value domains (e.g., autonomous driving, robotics, voice assistants, medical, finance, education). This paper revisits the argument that "common sense AI," even when exhibiting below human-level IQ, can address important tasks—such as SAE Level 5 Full Self-Driving (FSD), robotics, and voice assistance—if integrated with sufficiently resolved philosophical foundations for representative challenges (e.g., Chinese Room, Frame Problem, symbol grounding, alignment/paperclip) [Latapie, 2025a]. Our central list of philosophical concerns is open-ended rather than final or exhaustive; nonetheless, achieving robust starting solutions in these illustrative areas shows that smaller-scale AI systems, leveraging orders of magnitude less data and compute while reaching a minimal but sufficient IQ level[1], can reliably support a broad spectrum of real-world applications. Meanwhile, we must also consider legitimate concerns about unaligned ASI (Artificial Superintelligence) that may arise if capabilities progress without an adequately rigorous, validated foundation—for instance, deceptive or hallucinatory outputs. We suggest deferring efforts to achieve ever higher IQ AI until further solutions to deeper moral and safety challenges are validated. Non-automated or non-agentic use cases that leverage partially unaligned AI can continue to be beneficial if labeled appropriately to prevent overtrust. Overall, this phased roadmap suggests AI, rooted in the well-defined frameworks of "common sense AI" plus evolving philosophical solutions, can yield immediate benefits while reducing resource consumption—and do so without risking unaligned ASI autonomy.

***K*eywords** Artificial Intelligence, Common Sense, Minimal Prior Knowledge, Contextual Reasoning, Chinese Room, Frame Problem, Symbol Grounding, Alignment, Paperclip, Unaligned ASI, SAE Level 5 FSD, Robotics, Voice Assistants, Philosophical Foundations

1 Introduction and Rationale

1.1 Scaling Alone Shown Insufficient for High-Value Domains

High-profile companies once championed the idea that ever-expanding datasets, model sizes, and compute would yield robust AI. However, recent developments illustrate that such a purely scaling-centric approach has encountered practical and economic strain:

- **Apple** halted AI-based news curation after reliability shortfalls harmed user trust [Press, 2025].
- **Amazon** struggles with "hallucinations" in Alexa generative pipelines, hampering ROI [Brodkin, 2025].
- **OpenAI** CEO Sam Altman publicly moderates AGI hype, citing real cost-performance barriers [Magazine, 2025].

With disclaimers and user mistrust persisting, "just scale it more" alone fails to deliver robust contextual capability.

[1]We use the term IQ in this document as a stand-in for cognitive ability. We are not opining on whether it is the best metric or if current IQ tests and benchmarks are valid or not valid

1.2 Why Minimal-But-Sufficient IQ AI Helps

Modern AI systems—like virtually all human technologies—can show superhuman capacity in certain narrow aspects (e.g., trillions of operations per second, specialized benchmark performance). Yet these feats, by themselves, do not guarantee alignment, reliability, or beneficial outcomes across the varied spectrum of real-world use cases. Many models surpass average human performance in specialized tasks yet can still produce surprising misalignment or illusions, affecting everything from autonomous deployments to customer service chatbots. Drawing from "Common Sense Is All You Need" [Latapie, 2025a], we posit that even if an AI remains below human-level IQ, it can still achieve robust operational success by adhering to a well-defined common sense AI paradigm. Meanwhile, addressing key philosophical challenges—such as the Chinese Room, Frame Problem, symbol grounding, and alignment/paperclip—further ensures reliability and trustworthiness. Moreover, operating with far fewer computational resources reduces the environmental footprint while still fulfilling major commercial and societal needs.

1.3 Aim and Outline

We propose a phased framework:

1. Deploy below human-level IQ AI—anchored in common sense AI plus robust solutions to key philosophical validations—for high-value tasks such as FSD, robotics, and voice.
2. Acknowledge legitimate worries regarding unaligned ASI (i.e., potential superintelligence that may lack adequate alignment) by deferring increases in IQ until deeper moral and safety frameworks are validated.
3. Leverage existing unaligned AI in non-mission-critical use cases where hallucinations and current accuracy rates are acceptable. Clear and concise labeling is crucial to prevent misunderstandings of capability and overtrust.

Sections detail how reliance on scaling alone is slowing progress (Section 2), how common sense AI plus resolutions to key philosophical challenges is a path to beneficial AI capable of solving a wide array of important use cases (Sections 3-4), and how increases in IQ must solve additional hurdles to avoid unaligned ASI (Section 5).

2 Market Realities: Why Scaling Alone Can Stall

2.1 High Costs, Limited Returns for Overly Resource Intensive Models

Even though AI already outperforms humans in certain benchmarks, problems with hallucinations and reliability continue to hinder deployment of high-value real-world solutions. Apple, Amazon, and OpenAI collectively illustrate that scaling data and model size alone cannot surmount hallucination or alignment pitfalls. Another drawback of the current scaling approach is the massive infrastructure costs and environmental impacts.

2.2 User Trust Eroded by Hallucinatory, Misleading Outputs

Absent a robust "common sense AI" foundation, advanced systems often produce manipulative or nonsensical content. Such deployments diminish brand trust and stall monetization, indicating a need for new directions.

3 Common Sense AI with Minimal-But-Sufficient IQ, Supported by Key Philosophical Principles

3.1 Foundations of Common Sense AI

Common sense AI is founded on **Minimal Prior Knowledge (MPK)**—the ability of an AI system to infer, adapt, and respond meaningfully to its environment without relying on vast pre-trained datasets. This approach emphasizes: - **Contextual Sensitivity**: The system dynamically prioritizes relevant information and dismisses irrelevant details in real-time interactions. - **Robust Adaptation**: By interacting with its environment, the system demonstrates resilience and reasoning in novel or ambiguous situations. These attributes ensure that common sense AI avoids brittle, heuristic-driven behaviors, creating systems that are flexible, transparent, and contextually aware [Latapie, 2025a,b].

3.2 Philosophical Challenges Supporting Common Sense AI

To achieve genuine common sense, AI must address critical philosophical challenges that often expose the limitations of scaled, heuristic-based systems. These challenges include:

- **Chinese Room Argument** [Turing, 1950]: Avoids purely syntactic processing by ensuring that AI systems develop genuine semantic understanding, enabling meaningful interactions rather than surface-level mimicry. - **Frame Problem** [Chollet, 2019]: Tackles the complexity of dynamically identifying relevant factors in a changing environment, avoiding computational paralysis or trivial reasoning errors. - **Symbol Grounding**: Establishes consistent mappings between internal representations (symbols or tokens) and real-world entities, ensuring that system outputs remain logically grounded and interpretable. - **Alignment and the Paperclip Problem** [Bostrom, 2014, Latapie, 2025c]: Embeds moral and ethical constraints into AI systems, ensuring they align with human values while avoiding unintended optimization traps.

By integrating solutions to these challenges, common sense AI provides not only practical utility but also safeguards against issues like manipulation, hallucinations, or misaligned outputs.

3.3 Advantages of Common Sense AI with Minimal-But-Sufficient IQ

Deploying AI systems rooted in common sense principles offers significant advantages, especially when intelligence is calibrated to a minimal-but-sufficient level:

1. **Sustainability**: Smaller, task-focused models consume less computational energy, reducing environmental and financial costs without sacrificing effectiveness. 2. **Improved Reliability**: Grounded reasoning and philosophical solutions reduce the occurrence of nonsensical or manipulative outputs, increasing trust in real-world applications. 3. **Broad Applicability**: Systems with foundational common sense can address diverse high-value use cases, from autonomous systems to interactive tools, without relying on massive data or compute resources. 4. **Contextual Robustness**: The system's ability to dynamically adapt and reason based on minimal but sufficient information ensures reliability across novel or edge-case scenarios.

4 Use Cases: FSD, Robotics, Assistants, and Limited-Application AI

4.1 SAE Level 5 Full Self-Driving (FSD)

Why Minimal-But-Sufficient IQ Works When built on the principles of common sense AI, systems with below human-level IQ can effectively interpret road semantics, track real-time traffic, and adhere to safety constraints. By employing dynamic relevance filters, the Frame Problem is mitigated. These systems retain access to advanced computational capabilities and specialized tools, enabling superior analytical performance while capping their IQ to focus on the specific task at hand. Such systems can be rigorously validated and offer more predictable behavior.

Environmental and Resource Benefits Reduced computational demands translate to lower infrastructure costs and energy consumption. This enables large-scale deployment in vehicles without the need for expensive, high-powered onboard supercomputing hardware, achieving greater scalability and sustainability.

4.2 Robotics: Household, Warehouse, Industrial

Context Over Encyclopedic Knowledge Robots equipped with below human-level intelligence, supported by access to advanced tools and enhanced by environment-aware reasoning, can perform tasks such as object manipulation, obstacle navigation, and safe interaction with humans. By leveraging symbol grounding and achieving minimal-but-sufficient IQ levels, these systems can deliver high performance in specific tasks without the need for vast computational resources or encyclopedic knowledge.

Stability and Trust Defining problem scopes and use cases for robotics ensures stability and user trust. While generality remains feasible within defined domains, these systems avoid unnecessary overreach. Disclaimers for potential edge cases help maintain reliability in deployment.

4.3 Assistants: Home, Personal, and Educational Applications

Broadly Applicable Assistants Common sense AI principles underpin a broad range of assistant applications, including home assistants, personal productivity tools, and educational companions. These systems, whether voice-based or otherwise, operate with minimal-but-sufficient IQ while retaining access to specialized tools that enhance their utility. This approach ensures reliable support across a variety of domains without requiring disclaimers or overgeneralized claims.

Avoiding Hallucinations Assistants designed with a core of common sense AI and guided by philosophical principles (e.g., Chinese Room, alignment) avoid producing hallucinations or nonsensical outputs. Instead, they deliver coherent, meaningful results tailored to specific user needs.

4.4 Labeling Partially-Aligned AI for Limited Use Cases

AI systems that are not fully aligned or designed for general-purpose applications can still excel in limited domains such as generative AI, data analytics, summarization, or content recommendation. By explicitly labeling these systems as task-specific tools with clearly defined limitations, users are less likely to overtrust their capabilities. This approach allows for responsible and effective deployment, even when accuracy and alignment do not meet the highest standards required for broader use cases.

5 Above-Human Intelligence and Avoiding Unaligned ASI

5.1 Ongoing Limitations of Advanced Models

Many foundation models demonstrate above-human performance on specific benchmarks but continue to exhibit hallucinations, contextual misalignment, and failures in reasoning. These issues undermine trust and reliability across use cases, ranging from creative tasks to critical systems. Without a foundation rooted in common sense and solutions to key philosophical challenges, these models risk evolving into unaligned ASI, with potentially harmful consequences.

5.2 How to Prevent Unaligned ASI

The continued development of higher IQ AI systems should proceed with extreme caution, particularly when such systems lack alignment, common sense, or solutions to key moral and safety challenges. These systems must be deployed only in contexts where their known limitations—such as hallucinations or misaligned behaviors—will not result in harm or unintended consequences.

In all cases, deployment must be accompanied by full disclosure of capabilities, limitations, and risks, ensuring informed consent and transparency for users. Systems should not be used in high-stakes or morally critical applications without validated frameworks for alignment and safety. By adopting a "both eyes wide open" approach, scaling higher IQ systems can be explored responsibly, mitigating the risks of unaligned ASI while leveraging their potential benefits in safer, limited-use scenarios.

6 Conclusion and Future Directions

6.1 A More Grounded Path Forward

The limitations of scaling data, models, and compute alone have become apparent, particularly regarding trust, reliability, and alignment. As an alternative, we propose a roadmap that integrates:

- Developing and validating a "common sense AI" framework capable of handling both non-verbal and verbal use cases, as future work to provide an existence proof,
- Addressing key philosophical challenges (e.g., Chinese Room, Frame Problem, symbol grounding, alignment/paperclip) as foundational components of robust AI,
- Deploying below human-level IQ AI systems augmented with super-human compute and tools as needed, ensuring they meet the minimal-but-sufficient IQ criteria for their intended use cases while adhering to foundational principles.

6.2 Practical Steps for Industry

Adopt Grounded Frameworks for Targeted Applications Rather than relying on scaling as a default strategy, industry leaders can explore solutions grounded in emerging frameworks like common sense AI. While this framework is not yet fully established, early steps toward integration—such as targeting specific domains with minimal-but-sufficient IQ systems that adhere to robust foundations—offer a path forward for safer and more reliable AI.

Deploy Specialized Systems for Defined Use Cases AI systems with below human-level IQ, augmented by advanced computational tools, can deliver super-human performance in specific tasks while minimizing risks associated with

artificial super intelligence. These systems provide an opportunity to address real-world problems with efficiency and precision, without the complexity or unpredictability of unconstrained and unaligned high-IQ systems.

6.3 Safe Advancements Toward Greater Capabilities

The possibility of leveraging such systems to act as gatekeepers for the development of ASI is an intriguing direction for future exploration. However, we are far from recommending this approach at this stage. For now, caution remains essential when deploying AI in high-stakes scenarios. Systems should be accompanied by clear disclosures, rigorous validation, and transparency regarding capabilities and limitations to mitigate risks of unaligned or deceptive behaviors.

6.4 Final Outlook

Scaling alone cannot address the critical challenges of alignment, trust, and environmental sustainability. A grounded approach—centered on emerging frameworks like common sense AI, philosophical rigor, and targeted deployment—offers a more promising path forward. While much work remains to establish and validate these principles, the proposed roadmap prioritizes immediate, practical benefits while reducing risks. Over time, this approach could pave the way for responsibly scaling AI capabilities, ensuring that advancements align with human values and serve humanity's broader interests without risking unaligned ASI.

References

Hugo Latapie. Common sense is all you need. *arXiv preprint*, 2025a.

Associated Press. Apple pulls error-prone ai-generated news summaries in its beta iphone software. *AP News*, January 21 2025. https://apnews.com/article/apple-ai-news-hallucinations-iphone-6b37a11b9cdd0e100c299e922d58b530.

Jon Brodkin. Amazon must solve hallucination problem before launching ai-enabled alexa. *Ars Technica*, January 14 2025. https://arstechnica.com/ai/2025/01/amazon-must-solve-hallucination-problem-before-launching-ai-enabled-alexa/.

Analytics India Magazine. Sam altman: "agi is still a long way off. let's focus on responsible ai development.". Posted on X (formerly Twitter), January 20 2025. https://x.com/Analyticsindiam/status/1881298296695607688.

Hugo Latapie. Towards a litmus test for common sense: Axiomatic foundations for beneficial ai. *arXiv preprint*, 2025b.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

François Chollet. On the measure of intelligence. *arXiv preprint*, arXiv:1911.01547, 2019.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Hugo Latapie. Towards ethical ai. *arXiv preprint*, 2025c.