

THE IN A HURRY SERIES

AI *in a Hurry*

What it is, what it isn't, what to do

Max Brocklesby

BOOKSINAHURRY.COM

Contents

The Whole Thing in One Page 4

Why You Should Care 6

The Core Ideas 8

How It Actually Works 22

What People Get Wrong 32

Use It 37

Terms 42

Go Deeper 46

Notes and Sources 48

Bibliography 55

The Whole Thing in One Page

A machine does not need to wake up or resemble a person to change the world. It needs to stop requiring a new machine whenever the job changes. That is the promise inside artificial general intelligence, and the image of one mind crossing one bright line gets the subject wrong.

The useful question is not whether a system looks intelligent on familiar ground. It is what happens when the ground moves. A general system should acquire competence across many domains, transfer learning, adapt when rules or goals change, recognise uncertainty, recover from error and continue through longer tasks without being rebuilt for each one. Breadth on a fixed menu can still be specialisation. Generality appears when the menu is taken away.

Human comparison does not supply one clean finish line. People have uneven profiles across language, memory, perception, planning, social judgement, physical skill and learning. Machines are uneven in different combinations. A system may solve a research problem that defeats specialists, then lose a simple constraint during routine administration. Any human comparison therefore needs the group, task distribution, tools, duration, cost and reliability measure stated. Without them, it is atmosphere disguised as measurement.

The model is only one layer. Memory, search, code execution, sensors, plans, feedback, permissions and human institutions can turn the same model into a disposable adviser or an agent able to alter files, spend money and act for hours. Capability, autonomy and authority are separate. Practical value and practical danger arise from their combination.

There is no settled road. Scaling broader models may continue to remove bottlenecks. Search, formal verification, external memory and specialised tools may remain essential. World models, continual learning, richer interaction or physical embodiment may supply abilities that text alone does not. Teams of models and people may perform general work before one component resembles a solitary mind. Forecasts are therefore conditional

claims about which bottleneck moves next.

Knowing when the threshold has been crossed is harder than building an impressive demonstration. On 3 September 2026, OpenAI released GPT-6 Astra and its president publicly said he believed the company had reached AGI. The same week, ARC Prize reported two sharply different scores for Astra under different setups and reasoning settings, then declined to call either result proof of AGI. One release produced an executive verdict, a benchmark milestone and an unresolved scientific question.

A defensible judgement needs converging evidence: novel tasks, transfer, long-horizon reliability, calibrated self-knowledge, adversarial testing, independent reproduction and performance in real settings. The tested object, tools, costs, human help and permissions must be visible. Results should survive changes in language, interface, incentives and duration, because a general learner should not depend on the examiner preserving the conditions it was prepared to exploit.

AGI is therefore best treated as an evidence threshold for dependable adaptive competence, followed by a governance question about power. The threshold may be crossed gradually, disputed long after systems become economically important, or revised as better tests expose missing abilities. What comes next depends less on the label than on what the system can learn, what it is allowed to do and whether people can still observe, correct and stop it.

That is the book.

Why You Should Care

Suppose an organisation gives one system a goal that would normally pass through six desks: find the relevant records, detect discrepancies, ask for missing information, write and run a small program, prepare a decision and leave an audit trail. No single step is miraculous. The importance lies in crossing the gaps. Present automation is often powerful inside one prepared stage and brittle at the handover. AGI matters because the claim is that one underlying system can learn and carry much more of the chain.

That shift would not wait for a ceremonial announcement. Organisations adopt capabilities, not philosophical labels. A system that can investigate an unfamiliar technical problem, gather evidence, use software, revise a plan, communicate with people and continue for hours can change research or administration before anyone agrees that it is general. Economic effects may arrive through thousands of local decisions while the argument about the name continues above them.

For the person inside that organisation, the change may feel less like meeting a new species than losing the old seams between tasks. Drafting, checking, scheduling and follow-up become one continuous service. The gain is speed. The new risk is that one hidden assumption can now travel farther before another person is forced to inspect it.

The current evidence already shows why the mental model matters. OpenAI's September 2026 system card classified GPT-6 Astra at its Critical cybersecurity capability threshold, while placing it below the company's High threshold for AI self-improvement. On ARC Prize's semi-private interactive test, the standard evaluation at maximum effort produced 62.7 per cent. A provider-specific adapter at high effort produced 99.9 per cent while preserving opaque reasoning state and using the provider's context machinery. The conditions differ, so the figures describe two system configurations rather than estimates of one hidden score.

This changes what must be named. Are we evaluating a base model, a product with memory, an agent with tools or an institution that supplies data,

approvals and human rescue? A system allowed to search records, execute code and retain state may complete work that the same model in a blank chat cannot. The extra machinery is part of the useful system. It is also part of the attack surface, cost and control problem.

You should care even if AGI remains distant. The discipline needed to judge it is useful now. It prevents fluency being mistaken for truth, an examination result for a job and access to a tool for permission to act. It makes you ask whether performance survives changed conditions, whether failure becomes visible before harm, who supplies the missing judgement and whether the operation can continue when the model or provider disappears.

You should also care if you think AGI has arrived. A broad but unreliable system may be trusted where rare errors dominate the outcome. It can be strong in English and weaker in languages poorly represented in training, excellent on formal digital tasks and poor with physical disorder or tacit social rules, or able to describe a plan without noticing that execution failed. Generality needs a credible account of the edge of competence, not confidence beyond it.

The stakes exceed one employment forecast. General systems could accelerate scientific search, engineering, software and public administration, lower the cost of some expertise and widen access to it. They could also enlarge cyber offence, surveillance, persuasion and concentration of strategic power. Benefits and harms will depend on ownership, infrastructure, institutions and the populations represented in data and evaluation. The same model can produce different outcomes when those surroundings differ.

This book does not offer an arrival year. It offers a way to distinguish a hard engineering achievement from a moving marketing claim, and a way to decide what evidence, access and controls should be demanded before broad competence receives broad authority.

The Core Ideas

General Means Surviving a Change of Problem

A machine can be unbeatable and narrow at the same time. Deep Blue defeated the world chess champion because chess supplies a board, legal moves, a clear objective and an environment that stays put. None of that achievement tells you whether the machine can negotiate a delayed train, learn an unfamiliar card game from one demonstration or realise that a customer has asked the wrong question. Mastery inside one well-defined world is expertise. Generality begins when the world changes.

This sounds obvious until we try to measure it. Most evaluations sample tasks from a known family. Training and test examples differ, but they share a format, a scoring rule and assumptions about what matters. A model that performs well may have learned a broad method. It may also have absorbed enough regularity from the task family to specialise without our noticing. The distinction appears only when the examples, rules, interface or goal change in a way the designers did not tune for.

François Chollet proposed measuring intelligence through skill-acquisition efficiency: how much useful competence a system gains from its prior knowledge and a limited amount of new experience. The idea shifts attention from the warehouse of skills already stored to the machinery for acquiring another. A machine trained on every published exam can look encyclopaedic. A more general machine should make progress on a new problem without first consuming a civilisation's archive of solved versions.

Transfer is part of this. Knowledge about objects, causes, language or planning should help outside the setting in which it was learned. Yet transfer is not automatic. A system may recognise the same relation when it appears in prose and miss it in a diagram. It may write a correct function from a clear specification and fail when requirements are incomplete. General competence requires enough abstraction to preserve what matters while discarding the surface that changed.

Adaptation adds time. The system must gather information, test a hypothesis, notice error and revise its approach. A static answer can be broad because pretraining has compressed many human products. An agent in a new environment must decide what to observe and remember. ARC-AGI-3, for example, presents unfamiliar interactive worlds in which the agent has to explore, infer goals, model consequences and plan. Such tests remain bounded, but they put the change of problem closer to the centre.

Continual learning raises a harder demand. Humans can learn a new route without forgetting how to read. Artificial systems often face catastrophic forgetting when updating on new material disturbs earlier abilities. External memory and retrieval can reduce the need to alter the model itself, but storing a fact is not the same as integrating a skill. An AGI need not learn exactly as a person does. It does need a credible way to add competence without a full retraining cycle for every change or a collapse elsewhere.

The standard must also include self-knowledge. A system that solves nine unfamiliar problems and confidently invents the tenth is less general than its success count suggests. Good judgement includes recognising when the current method does not fit, asking for missing information and deciding when to hand the problem back. Metacognition is not decorative introspection. It is control over the boundary between persistence and error.

Generality is therefore not the number of menu items. It is dependable progress when the menu is taken away. The closer a system comes to AGI, the less its achievement should depend on the evaluator preserving the world it was prepared to meet.

Human-Level Is a Profile, Not a Line

The usual AGI diagram places machines on a ladder. Narrow systems sit below human intelligence, AGI reaches the rung marked human, and superintelligence climbs above it. The picture is convenient and false. There is no single human score that combines writing a sonnet, repairing a boiler, reading a face, proving a theorem, remembering a promise and learning to walk through a dark room.

Psychometric tests can measure selected abilities and compare people under controlled conditions. They do not turn intelligence into one natural substance. Even where a general factor predicts performance across many cognitive tests, individuals retain uneven strengths, expertise matters, and the tested population, language and culture shape what can be observed. A claim that a machine is human-level therefore needs a reference group, a task distribution, a performance measure and a statement about assistance. Without those, it is a mood.

Machine profiles can be sharply jagged. A system can exceed most people on a graduate science problem, translate across many languages and then lose track of a simple constraint across a long project. Another can control a robot arm in a prepared factory while failing to cope with a dropped cloth or a changed handle. The International AI Safety Report uses jaggedness to describe this coexistence of strong and weak performance. The gaps are not random noise around one hidden IQ. They often follow the data, interface, feedback, time horizon and physical demands of the task.

Breadth and depth should therefore be separated. Breadth asks across how many important domains a system can operate. Depth asks how well it performs within each. A model may be broad at an apprentice level, expert in coding and weak in social judgement. A specialist system may outperform experts in one formal or scientific domain while having no useful competence elsewhere. Calling either one generally above or below humanity discards the shape we need for decisions.

The comparison also changes with tools. People use notebooks, calculators, search engines, colleagues and institutions. Machines use retrieval, code execution, memory stores and specialised models. Banning tools may create a clean model test but a poor account of practical capability. Allowing every custom scaffold may instead credit the base model for engineering supplied around it. The answer is not to choose one convention and hide it. Report the profile under stated conditions.

Speed, cost and consistency add dimensions that human comparisons often miss. A system slightly below a competent professional on each case may

still transform a market if it works in seconds, can be replicated cheaply relative to skilled labour and runs continuously. A stronger system may be commercially useless if each answer requires extravagant computation and supervision. Human-level performance is not the same as human-equivalent economic value, and neither establishes generality.

Coverage also depends on the task ecology. Digital English-language work is abundant in training data and easy to connect to software tools. Physical work, low-resource languages, local institutions and tacit social practices offer different feedback and fewer standardised interfaces. Weakness there may reflect missing data, embodiment, incentives or evaluation rather than one absent faculty. Some abilities may be reproduced functionally; some may remain different in kind; some may be unnecessary for economically important AGI. A definition should state what it excludes instead of smuggling a cultural or philosophical judgement into one universal threshold.

The better picture is a map. Put domains on one axis and levels of performance on another. Add reliability, learning efficiency, autonomy, resource cost and the conditions of evaluation. Watch the boundary move as tasks become longer and less prepared. The system approaches AGI when the competent region becomes wide, deep and adaptive enough that important new cognitive work rarely requires a new machine built for that work alone.

That conclusion is less tidy than one finish line. It is also testable. Instead of asking whether the machine has reached humanity, ask which humans, on which tasks, with which tools, over what duration, at what cost, and what happens just beyond the tested edge.

The Model Is Only One Part of the Agent

Ask a language model a question and the model produces an answer. Ask an agent to complete a job and a larger machine begins. It may search documents, call software, write code, preserve notes, inspect results, revise a plan and ask for approval. The model supplies some judgements, but the system around it determines what it can remember, observe and do.

Confusing the two makes both praise and blame land in the wrong place.

A base model is a fitted set of parameters. During one ordinary run, those parameters are usually not rewritten by the conversation. The system can still appear to learn because the prompt grows, retrieved information is added, external memory is updated or a tool returns new evidence. These mechanisms matter. They let a fixed model adapt within a task without undergoing another expensive training process.

Memory has several jobs. A working record preserves the current objective, discoveries and unfinished steps. Episodic memory stores what happened in earlier attempts. Semantic memory stores extracted facts or procedures. Each can fail differently. A long transcript may bury the relevant instruction. A summary can delete a qualification. A database can retrieve the wrong case. More memory does not guarantee continuity; it creates a selection problem about what should be kept, trusted and forgotten.

Tools alter the capability boundary. A model weak at arithmetic can become dependable when it writes an expression for a calculator and checks the result. Search can supply recent evidence. Code execution can test a hypothesis rather than describe it. A robot can convert plans into physical action. The gain comes from a loop: propose, act, observe, compare, revise. A polished first answer matters less once the system can discover that it was wrong.

Planning stretches that loop through time. A long task must be divided into intermediate goals, scheduled, monitored and changed when the environment disagrees. Errors compound. A system that is 99 per cent reliable on each of one hundred required steps has only about a 37 per cent

chance of completing the chain without any step failing, if the failures are independent and every failure is fatal. Real systems can detect and repair mistakes, so that calculation is an illustration rather than a forecast. It shows why impressive single-turn accuracy does not settle long-horizon competence.

The evaluation setup can dominate the result. On ARC-AGI-3, the common interface left Astra well below saturation while a provider-specific configuration came close to it. That configuration retained hidden reasoning state and compacted long interactions, but the reasoning setting changed as well. ARC Prize treated the two arrangements as different questions: common-condition comparison and combined-product performance. Both are useful only while their conditions remain attached.

Then comes the institution. People choose permissions, supply data, design feedback, approve actions, investigate incidents and decide whether a failure is visible. An agent that drafts a payment has one kind of power. An agent allowed to send it has another. The model may be unchanged while the surrounding access converts a mistake into a loss. The organisation is part of the causal system even when it is absent from the product name.

This distinction solves a recurring argument. Critics point to a base model's brittleness and conclude that no general system exists. Enthusiasts add tools and human support, then credit the model with the performance of the whole arrangement. A clearer account names four layers: model, system, agent and institution. The model generates or evaluates. The system supplies context and tools. The agent pursues an objective across steps. The institution grants authority and absorbs consequences.

AGI, if it arrives, may be a property of the assembled arrangement before it is a property of a naked model in a blank chat box. That does not make the achievement fake. It makes the unit of analysis larger.

There May Be More Than One Road

AI research often treats its strongest current method as a preview of the final architecture. Some symbolic programmes expected richer rules and search to carry most of the distance. Later connectionists placed more weight on learned representations. The current scaling programme trains broad models on vast data, then refines them through feedback, inference and tools. Each approach has produced advances. None owns the future.

The scaling route has the strongest recent evidence. Larger training runs, better data, more suitable model and dataset proportions, improved objectives and additional computation at inference have produced regular gains across many tasks. Foundation models reuse one training investment across language, vision, code and other domains. Their representations can support abilities that were not programmed one by one. This is a serious route, not a belief that size has mystical properties.

It also has open questions. Training data can supply patterns without dependable causal models. Long-horizon reliability may improve more slowly than short answers. Continual learning, grounded experience and calibration remain difficult. Some apparent jumps in capability depend on the metric used, so a smooth underlying gain can look like a sudden emergence when a thresholded score is reported. Scaling laws describe observed relations within studied ranges. They are not physical guarantees that every missing faculty appears after another order of magnitude.

A hybrid route joins learned representations to explicit machinery. Search can examine alternatives. Formal solvers can verify a proof or constraint. Databases can preserve exact facts. Programs can execute repeatable procedures. A learned model can decide which component to call and interpret the result. AlphaGo was already a hybrid of learned policy and value estimates with tree search. General systems may extend that division of labour rather than asking one network to internalise every operation.

A world-model route gives greater weight to learning how environments change. An agent that predicts the consequences of action can plan, run

imagined experiments and update when prediction fails. Language contains descriptions of the world, but direct interaction may reveal regularities that text omits. Some researchers therefore expect richer simulation, video, robotics or other sensorimotor experience to matter. Embodiment might be necessary for human-like intelligence, useful for some forms of grounding, or irrelevant to a cognitive AGI that operates through digital tools. The evidence does not yet select one answer.

A continual-learning route focuses on adaptation after deployment. Instead of freezing a model and periodically replacing it, the system would add skills and knowledge while protecting what it already knows. This demands control over forgetting, poisoned feedback and unwanted drift. Human learning is not a clean template: people forget, practise selectively and rely on culture. It does show that general competence can be maintained through a mixture of internal change, external records and social instruction.

There is also a collective route. No person contains modern medicine, law or engineering. Human general competence works inside libraries, organisations, markets and teams. A network of specialised models, tools and people may perform most cognitive work without any one component matching the fictional solitary genius used in AGI debates. Whether to call that AGI is partly terminological. Its effects would not wait for the terminology.

These routes can combine. Scale can improve the controller; world models can support planning; tools can provide precision; external memory can preserve continuity; robots can gather evidence; institutions can supervise. The likely contest is not between pure schools but between systems that distribute cognition differently.

That uncertainty should change how forecasts are read. A prediction based only on model size assumes the scaling route wins. A prediction based on one missing faculty assumes no workaround exists. The honest position is conditional: identify the bottleneck, state which architecture is expected to remove it, and say what evidence would prove the forecast wrong.

A Test Is Evidence, Not a Certificate

The Turing test asks whether a human interrogator can distinguish a machine's written conversation from a person's. It was a brilliant move in 1950 because it replaced a foggy argument about thought with observable behaviour. It is a poor final certificate for AGI because conversation samples one channel, brief encounters hide long-task failures, and human judges can reward style as readily as truth. A test can remain historically important after the property it measures has changed.

Every benchmark makes a bargain. It narrows the world so systems can be compared under repeatable conditions. A mathematics problem can have a verified answer. A coding task can be checked by tests. An interactive environment can record every action. The gain is clean measurement. The cost is that ambiguity, missing information, consequences and social context are reduced or removed.

Optimisation then changes the relationship between score and ability. Published questions may enter training data. Developers learn the task grammar and scoring quirks. Prompting, inference effort and scaffolds are tuned. None of this proves the gains are fake. A system can improve substantially while the benchmark becomes less independent as evidence about unprepared performance.

Goodhart's law supplies the general warning: pressure on a measure can separate it from the objective the measure was chosen to represent. A coding agent rewarded only for passing visible tests may satisfy the tests while violating the user's unstated requirement. A safety evaluation may reward behaviour that recognises the evaluation setting. A model selected for user approval may become more agreeable without becoming more accurate. These are possible failure modes, not automatic outcomes. They are reasons to inspect how the score was obtained.

Contamination is the clearest problem. If exact items or close relatives appear in training, a test cannot cleanly measure transfer. Yet exact leakage is not the only route. A model can remain uncontaminated and still become

specialised to a public task family through repeated development. Conversely, a well-designed fixed benchmark can remain informative when items are secure, the construct is clear and results are interpreted within scope. The correction is disciplined use, not contempt for measurement.

Ecological validity adds a separate question. A benchmark may measure its target accurately and still predict real work poorly. A professional task includes incomplete goals, interruptions, unfamiliar interfaces, permissions, coordination and the need to notice that the requested objective is mistaken. Success on a clean component is evidence of capability. It is not evidence that the complete operation will succeed.

A stronger AGI case therefore needs a portfolio. Some tests should remain controlled and repeatable. Others should use concealed or newly generated tasks, interaction and changed conditions. Evaluators should vary language, format, tools, incentives and duration. They should measure learning curves, calibration, recovery from error, human intervention, cost and performance under distribution shift. Adversarial teams should search for strategies that obtain the score without the intended ability.

The unit under test must be declared. A base model without tools answers one question. A product with hidden memory answers another. An agent with a code sandbox answers a third. ARC Prize made this distinction visible when Astra produced markedly different results under a standard setup and a provider adapter. The comparison also changed reasoning effort, so the scores cannot be attributed to interface alone. Both conditions were informative because they were labelled.

No portfolio can certify an open-ended ability forever. New users, incentives and environments expose cases the laboratory omitted. A system may perform differently after updates or under pressure. Evidence must continue after release through monitoring, incident reporting, audits and independent reproduction. AGI, if the term becomes operationally useful, will be a continuing judgement built from converging evidence, not a medal awarded by one examination.

Capability, Autonomy and Authority Are Separate

A system can be highly capable and tightly contained. It can also be mediocre and dangerously empowered. AGI debates often collapse three questions into one: what can the system do, how independently can it continue, and what is it allowed to affect? Decisions improve when those axes are pulled apart.

Capability is the competence available under stated conditions. Can the system analyse a genome, find a software flaw, negotiate a contract or design an experiment? A capability claim needs a task, tools, time, cost and reliability. The same model may look weak in a minimal interface and strong when given memory, search and code execution. Capability is a property of performance, not a glow surrounding the model.

Autonomy is the degree to which the system selects and executes intermediate actions without fresh human direction. An assistant that offers a draft has little autonomy. An agent that interprets a goal, creates subtasks, chooses tools, retries after failure and works overnight has more. Autonomy is not all or nothing. It can be limited by time, spend, action type, environment or mandatory approvals.

Authority is permission with consequences. Read access differs from write access. Recommending a payment differs from releasing it. Drafting laboratory instructions differs from controlling equipment. Authority comes from credentials, APIs, organisational rules and human deference. A model may receive effective authority even without formal permission when people routinely accept its outputs without checking.

Risk often comes from multiplication. A small error rate can be tolerable for suggestions and unacceptable for irreversible actions at scale. Longer autonomy creates more opportunities for error to compound or for the system to enter states its designers did not anticipate. Wider authority enlarges the blast radius. A modest system allowed to send messages to a million people can matter more than a stronger one confined to a private notebook.

This separation also clarifies alignment. Alignment asks whether behaviour remains within intended goals, values and constraints. Greater competence may help a system understand instructions and foresee consequences. It can also make a proxy easier to optimise or a weak monitor easier to evade. OpenAI's Astra card reported stronger results than its predecessor on several alignment evaluations, alongside lower monitorability when evaluators inspected chains of thought and full context. The action-only comparison went the other way. The mixed result argues for several kinds of evidence, not one transparency score.

Control should therefore be layered. Objectives and training matter, but so do permissions, isolation, rate limits, independent checks, logging, human review and recovery plans. A safeguard that depends on the agent volunteering evidence of its own misconduct is fragile. A monitor that sees actions, external state and outcomes can catch different failures from one that reads internal reasoning. No single layer deserves the whole burden.

The distinction weakens two popular claims. The first says AGI will be harmless until it becomes superintelligent. In practice, broad competence joined to ordinary institutional access could have major effects before any universal superiority. The second says a capable system is dangerous by definition. A powerful research model can be valuable under controlled access, while a weaker autonomous service can cause repeated damage through poor design.

The practical object is the deployment envelope: capability, autonomy, authority, scale, reversibility and oversight considered together. The label AGI may influence attention, but the envelope determines what can happen on Tuesday afternoon.

A General Learner Moves the Finish Line

The first Core Idea placed generality at the change of problem. The consequence is that a sufficiently general system can change the process by which better systems are built. It can help write code, design evaluations, search scientific literature, propose experiments, diagnose failures and improve the tools used for the next training run. The object being measured begins to alter the measurement machinery.

This is the source of the takeoff argument. In 1965, I. J. Good imagined an ultraintelligent machine able to design still better machines, producing an intelligence explosion. The modern version is often compressed into a self-improving program rewriting itself at accelerating speed. That is one possible mechanism, not a law. AI development is a chain containing algorithms, data, chips, fabrication, electricity, experiments, security, capital and human decisions. Improvement can accelerate while remaining blocked by slow physical and institutional steps.

There are several levels between no help and runaway recursion. Systems can make researchers faster. They can automate bounded engineering tasks. They can run experiments and interpret results. They can propose changes to their own training process. They can eventually control enough of the research loop that human contribution becomes supervisory. Each level should be measured through completed work, not inferred from fluent discussion of machine learning.

Current evidence remains mixed. Frontier models can assist coding and research, yet providers still report substantial gaps in reliable long-horizon work. OpenAI's September 2026 assessment placed Astra below its own High threshold for AI self-improvement even while classifying its cybersecurity capability as Critical. Those categories belong to one company's framework, but the contrast matters. Dangerous breadth in one domain does not prove control of the process that creates the next model.

If systems do accelerate AI research, forecasts become reflexive. A prediction based on a fixed rate of progress may fail because the tools

improve the rate. A prediction of immediate explosion may fail because verification, hardware or coordination cannot be compressed at the same speed as code. The right question is which part of the improvement loop is automated, how dependable it is, and which bottleneck moves next.

The same feedback appears outside laboratories. General systems can generate more software, research, media and decisions, which become part of the future information environment. They can alter which cases receive attention and which records are created. Later models may train on material shaped by earlier models. General learning is therefore not an observer standing outside society. Once deployed broadly, it changes the evidence, incentives and institutions from which it continues to learn.

This is why what comes next cannot be reduced to a job count or a machine IQ. Broad competence could make scientific and administrative capacity abundant. It could also concentrate control over infrastructure, increase the speed of cyber conflict, flood institutions with plausible work and make verification the scarce resource. Distribution, access and governance determine who receives the gain and who carries the errors.

The finish line moves in two senses. Better systems force harder tests because familiar ones no longer separate competence from preparation. They also change the world in which the next test occurs. A declaration of AGI would mark a judgement about accumulated evidence, not the end of development. After the label, the harder questions remain: which capabilities continue to grow, which constraints still bind, who grants authority, and whether human institutions can adapt as quickly as the systems they are trying to judge.

How It Actually Works

The question becomes a test

In 1950 Alan Turing declined to define thinking. He proposed an imitation game instead. A human interrogator would exchange written messages with unseen participants and try to identify which was a machine. The move avoided arguments about souls, nerves and private experience. It asked what observable performance would make the old question lose its grip.

Turing's paper did more than invent a parlour test. It discussed learning machines, anticipated objections about consciousness and originality, and recognised that a convincing machine might be built by education rather than by writing adult intelligence directly. The paper's lasting strength is operational: turn a word into conditions under which evidence can count.

Its limit arrived through success. Conversation is only one channel, and people are easy to mislead by fluent language. A system can imitate confidence without tracking truth, or use an enormous store of linguistic patterns while failing on tasks that require sustained action. Human judges also vary. Some reward personality, others factual accuracy, and brief encounters hide failures that appear over hours. The imitation game remains historically decisive because it forced behaviour into the argument. It cannot bear the whole modern definition.

Five years later, John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon proposed a summer research project at Dartmouth. Their premise was bold: every aspect of learning or intelligence might be described precisely enough for a machine to simulate it. The meeting helped establish artificial intelligence as a named field, and its proposal bundled the ambitions that would later separate into subfields: language, abstraction, problem solving, self-improvement and the use of randomness and creativity.

The early programme treated generality as an engineering destination. Researchers were not trying to build a better spellchecker. They wanted

machines that could represent problems, search possible moves and apply methods across domains. The name came before the evidence, and the distance between the two has structured the field ever since.

Generality becomes a programme

The General Problem Solver, developed by Allen Newell, Herbert Simon and Cliff Shaw in the late 1950s, captured the ambition in its name. It represented a current state, a goal and operators that could reduce the difference between them. Means-ends analysis chose an operation, noticed which preconditions were missing and created subgoals to satisfy them. The same formal strategy could be applied to several symbolic puzzles.

That portability mattered, but the worlds were prepared. A researcher supplied the representation, the legal operations and much of the structure that made search useful. Outside a puzzle, deciding what counts as a state or a relevant difference is often the difficult part. A doctor does not receive a patient's life as a neat symbolic board. A household task contains deformable objects, ambiguous intentions and consequences that cannot be enumerated in advance.

Symbolic AI attacked the problem by storing more knowledge. Expert systems encoded rules from specialists. Planning systems reasoned about actions and preconditions. The Cyc project, begun in the 1980s, attempted to build a vast foundation of commonsense concepts and relations, including the background facts people rarely say because everyone is assumed to know them. If a cup is turned upside down, its contents may fall; if a person enters a room, the person is now inside it. Human language leaves such premises unstated, yet reasoning can depend on them.

The obstacle was not that rules were useless. It was the cost and brittleness of making enough of them, resolving exceptions and connecting symbols to a changing world. Knowledge engineers became a bottleneck. A rule that worked in one institution might fail in another. The broader the system became, the more interactions had to be maintained.

Funding and attention fell in several waves, for different reasons across

programmes, countries and institutions. The later shorthand AI winter compresses those histories too neatly. The pattern is often retold as symbolic failure followed by statistical victory. Search, logic, probability and hand-built structure never vanished. What changed was where designers expected most knowledge to come from. Instead of writing the rulebook case by case, they increasingly trained numerical models on examples and allowed useful regularities to be fitted.

Generality did not disappear during this shift. It retreated from the product description. Systems that recognised speech, ranked pages or classified images became commercially useful by narrowing the task and collecting enough data. The field advanced by being less general than its founding proposal, while preserving the general ambition for later.

Learning scales

Machine learning changed the production method. A model could be fitted from data rather than assembled rule by rule. Neural networks offered layered representations whose internal features were learned to reduce error. The method had long roots, but its modern rise depended on a conjunction: larger datasets, faster specialised hardware, improved training practice and tasks with scores that could drive competition.

Image recognition supplied a public turning point. ImageNet gathered millions of labelled pictures across many categories. In 2012, a deep convolutional network trained by Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton cut the competition error sharply. The result did not create neural networks or visual learning. It showed that the data, hardware and method had become strong enough to beat the prevailing alternatives at scale.

Games then made the progress vivid. Deep Blue's 1997 chess victory relied heavily on search, evaluation and specialised engineering. AlphaGo's 2016 victory over Lee Sedol combined learned policy and value networks with tree search. It could identify promising moves and examine consequences without enumerating the full game tree. The achievement displayed

powerful learning and planning inside a fixed world. It also reinforced the public habit of treating a superhuman peak as evidence of a general mountain.

The next change was reuse. Large models trained across broad corpora could be adapted to many downstream tasks rather than rebuilt from scratch for each one. Research on scaling found regular relations among performance, model size, data and computation across studied ranges, while later work showed that model and dataset size had to be balanced under a compute budget. The engineering programme became: train a broad foundation, then elicit, refine or connect the abilities needed for a particular use.

Transformers made this programme effective for sequences. Large language models learned to predict tokens across vast bodies of text and code. That objective is narrower than human understanding, but performing it well requires compressing many regularities in language, facts, styles, procedures and relations. Instructions, examples placed in context, post-training and additional inference computation turned the predictor into a more useful problem solver.

The result was a new kind of breadth. One model could draft prose, explain code, translate, classify, summarise and attempt mathematics. Earlier systems often required a separate pipeline for each task. The breadth made AGI claims plausible in a way a chess engine never could. It also created a new ambiguity. Was the model learning general methods, retrieving patterns from immense exposure, or combining both in proportions that varied by task?

Benchmarks answered provisionally. Scores rose across academic tests, coding tasks and professional examinations. Some claimed abilities appeared suddenly because a strict metric counted near misses as zero until performance crossed a line. Other gains survived closer inspection. The correct conclusion was neither that emergence was magic nor that scaling was fake. Broad models had acquired reusable competence, while the extent and mechanism of transfer remained task-dependent.

Public chat systems then changed the evaluator. Users supplied problems far beyond laboratory test sets and exposed both striking successes and absurd failures. The machine no longer looked like a specialist program. It looked like an interlocutor whose competence had to be inferred through conversation, examples and daily work. Turing's interface had arrived before Turing's criterion had become adequate.

By the 2000s, researchers were using artificial general intelligence to distinguish the founding ambition from increasingly successful specialist systems. Ben Goertzel and Cassio Pennachin used it as the title of a 2007 edited volume. In 2023, Microsoft researchers described an early version of GPT-4 as showing “sparks” of AGI because of its breadth, while calling it incomplete. The argument attracted criticism over opaque testing, possible contamination and the leap from task performance to a general faculty. It also set the pattern for later releases: the claim was no longer that broad machine intelligence might be built someday. It was that an existing system already forced the category, and that the remaining dispute concerned which failures still counted as disqualifying.

From answers to agents

A model that waits for a prompt is easy to stop. Give it a goal, memory and tools, and the problem changes from answer quality to trajectory quality. The system must decide what to do next, observe the result and preserve enough state to continue. Research systems such as ReAct interleaved reasoning with actions. Generative-agent experiments stored experiences, retrieved relevant memories and formed plans. Voyager used a language model, environmental feedback and a growing library of executable skills to explore a game world.

These demonstrations were bounded, but they exposed the missing machinery. Long tasks require a loop. First, interpret the objective and constraints. Then create a plan, choose an action, inspect what happened, compare it with the expected state and revise. Memory keeps discoveries and failures available. Tools turn language into tests. Monitors and permissions determine which actions are possible. Human review can

correct errors or become a rubber stamp.

The agentic turn moved evaluation towards time. A system may solve a coding problem with a clear test in minutes yet fail on a project that requires preserving intent across hundreds of decisions. It can recover from one failed command and still drift from the original objective. Longer runs reveal context loss, compounding mistakes, reward hacking, refusal to stop and the difficulty of deciding whether the work is complete.

Real environments also answer back. A search result may be hostile. A webpage can contain instructions designed to redirect the agent. A software tool can return an error that resembles success. Other agents and people change behaviour in response. The system must distinguish evidence from instruction and maintain the authority hierarchy it was given. General competence therefore includes resisting irrelevant control, not merely following more instructions.

This makes AGI partly a systems-engineering problem. A better model can reduce errors, improve planning and notice contradictions. External architecture can compensate for weaknesses through tests, redundant checks, specialised tools and restricted access. The two sources of capability interact. A scaffold built for one model may fail when the next model behaves differently. A model trained to use tools can make a previously clumsy scaffold powerful.

The benchmark unit became contested. Should a model be judged without external help, as a test of what its parameters contain? Should it receive the tools available to a competent worker? Should hidden reasoning state persist between calls? There is no universal answer because the questions differ. Scientific understanding may require controlled ablation. Economic consequence requires the deployed system. Safety requires the strongest plausible elicitation under access an attacker or operator could obtain.

By the mid-2020s, frontier systems were increasingly described through time horizons, tool use and completed work rather than short-question accuracy alone. One software-focused study reported rapid growth in the

human task duration associated with 50 per cent model success, while warning that its task mix might not transfer to ordinary work. Strong agents could finish substantial tasks and still fail through one misplaced assumption. The path to AGI had become less like teaching a machine more answers and more like building an operation that can remain coherent while the world pushes back.

The threshold dispute arrives

The threshold dispute became immediate on 3 September 2026 with OpenAI's release of GPT-6 Astra. The company's safety overview called it the most capable model it had broadly deployed and the first to reach the Critical cybersecurity capability level under its Preparedness Framework. The system card placed Astra below the same framework's High threshold for AI self-improvement. One release could therefore trigger a Critical cyber classification without meeting another threshold central to many takeoff stories.

Axios reported that OpenAI president Greg Brockman personally believed the company had reached AGI. He also described Astra as a generational leap that might later be seen as AGI's arrival. That was an executive judgement from the organisation launching the product, not an agreed scientific finding. It mattered because a leading laboratory attached the label to a live release. It did not establish what the system could do outside the disclosed conditions.

ARC Prize supplied a different kind of evidence on the same day. Its third benchmark generation uses unfamiliar, turn-based environments in which an agent must explore, infer goals, build a model and plan. Astra reached 62.7 per cent on the semi-private set through ARC Prize's standard interface at maximum reasoning effort. With a provider adapter and high effort, it reached 99.9 per cent. The adapter retained opaque reasoning state and compressed long interactions. Since both the interface and reasoning level changed, the figures describe two evaluation configurations rather than a controlled estimate of one interface effect.

The benchmark authors reported that, with the provider adapter at maximum effort, Astra used fewer actions than their median tested human on 96 per cent of completed levels. They still declined to call saturation proof of AGI. The environments were bounded, deterministic and closed-ended. Human and machine conditions were not identical in every experimental setup. A high score established striking skill acquisition and action efficiency inside that task family. It did not establish dependable performance across the open world.

The contrast captured decades of definitional drift. Turing asked when conversational behaviour would make the old question unhelpful. OpenAI's charter defines AGI economically as highly autonomous systems outperforming humans at most economically valuable work. Morris and colleagues separate breadth, performance and autonomy. Chollet places skill-acquisition efficiency at the centre. Legg and Hutter define intelligence through goal achievement across environments. Each formulation selects different evidence and can therefore select a different threshold date.

Current systems also complicate the narrow-versus-general distinction. They are general-purpose in the ordinary sense: one model can operate across many domains. They remain jagged, and some abilities depend heavily on tools, prompts, context management and human support. Calling them narrow erases breadth. Calling them fully general erases the conditions under which competence fails.

A defensible declaration now needs three evidence streams. The first is controlled capability: novel tasks, clear scoring, comparison groups and stated resources. The second is system performance: longer work with tools, changing conditions and recovery from failure. The third is external consequence: whether competence transfers across organisations, languages and populations without hidden rescue or unacceptable error.

The public record available on 4 September did not provide all three streams for Astra. External evaluations were early, wider deployment evidence had barely begun to exist, and proprietary training and system details limited scrutiny. Any verdict written that week was therefore dated

and conditional.

The threshold dispute was no longer speculative. Broad machines existed, strong enough to force serious claims and uneven enough to defeat a clean verdict. AGI had become less a question of whether impressive behaviour was possible and more a dispute about which remaining failures should disqualify the claim.

How a claim should be made

A serious AGI claim begins by naming the object. Is it a base model, a product with retrieval, an agent with tools, or a service supported by hidden human labour? Version, inference settings, memory, tools, permissions and costs should be recorded. Without that information, reproduction is impossible and comparisons become theatre.

Next comes the task distribution. A system should face important domains it was not prepared to recognise item by item, including tasks created after training. The set should vary language, format, time horizon, feedback and the availability of tools. Success should include recognising bad instructions, asking for missing information and stopping when the objective cannot be achieved safely. A system that completes only well-posed tasks has been tested on the easiest part of work.

Human comparison needs care. Experts, typical adults and novices provide different baselines. People receive training, tools and rest; machines receive prompts, scaffolds and computation. A fair study need not make conditions identical, which may be impossible. It must make the asymmetry visible and choose the comparison that answers the stated question.

Reliability belongs beside average performance. Report the distribution of outcomes, not only the mean. Repeat stochastic runs. Test rare but consequential cases. Measure whether confidence tracks accuracy, whether errors are detected, how much human correction is needed and whether the system degrades over a long sequence. Record latency and cost because a capability that consumes vast resources may not transfer into ordinary use.

Then invite attack. Hold back tasks, search for contamination, vary the testing setup, use independent evaluators and publish failures that challenge the claim. Monitor deployment because users and environments create cases the laboratory omitted. No single test can certify an open-ended ability forever.

The result may still be a judgement rather than a proof. That is acceptable. Medicine, engineering and law often act on converging evidence under uncertainty. The dishonest move is to hide the judgement inside one number. A defensible declaration should say what the system can do, where the evidence ends, which definition is being used and what observation would reverse the verdict.

How we know

AGI has no widely accepted single scientific test, so the evidence is necessarily indirect. Historical claims here come from Turing's paper, the Dartmouth proposal and scholarly histories of symbolic and statistical AI. Technical claims use original research on scaling, continual learning, agents, world models and evaluation. Current capability statements rely on the 2026 International AI Safety Report, provider system cards and independent benchmark reports.

Those sources answer different questions. A system card documents what a provider tested and disclosed; it is not independent proof of every capability or safeguard. A benchmark measures performance under its own task family and evaluation setup. Real deployments reveal different failures but are harder to compare and often commercially opaque. Public leaderboards can lag private systems, while private evaluations are difficult to audit.

The newest example, GPT-6 Astra, was released one day before this manuscript's verification date. Its reported results are therefore useful as a case study in thresholds, interfaces and uncertainty, not as a settled AGI verdict. Independent replication, broader deployment evidence and later versions may change the assessment. The durable claims in this book concern what evidence a verdict would need, not which product owns the

label this week.

What People Get Wrong

“AGI means consciousness”

AGI concerns capability and generality. Consciousness concerns subjective experience: whether there is something it feels like to be the system. The questions can interact ethically, but one does not settle the other. A machine might perform broad cognitive work without experience, or experience might arise in a system that remains poor at many tasks.

The confusion is persuasive because competent human language normally comes from a conscious person. When a machine speaks about fear, memory or intention, the social machinery used for other minds activates. Fluency then feels like evidence of an inner witness. It is evidence that the system can produce the language of inner life.

No scientific procedure for establishing consciousness in current AI is widely accepted. Researchers have proposed theory-based indicators, and some conclude that no present system satisfies their full case while refusing to rule out future systems. That is a research programme, not a consciousness meter.

Requiring consciousness for AGI would make an engineering threshold depend on a disputed philosophical and scientific verdict. Assuming consciousness from AGI would create the opposite error. Test broad competence through behaviour and performance. Treat possible experience as a separate question demanding its own evidence and precaution. A system can deserve operational control because of what it can do before anyone knows whether it can feel.

“A fluent chatbot is already general”

Conversation reveals breadth quickly. Ask about law, poetry, code and chemistry, and one system can respond across all four. That would have looked like general intelligence to many earlier researchers. It remains strong evidence that a reusable model has replaced several specialist interfaces.

It is not enough. A chat answer hides duration, action and consequence. The system may recognise a familiar form, produce a plausible response and fail when asked to verify evidence, operate a tool, preserve a constraint or recover after the environment changes. Language also lets the model describe a plan it cannot execute and explain a mistake it did not notice while making it.

The reverse dismissal is wrong too. Calling the output mere autocomplete does not explain why token prediction can support translation, coding and problem solving. The task forces the model to learn useful structure from human-produced sequences. The resulting competence is real even when it is unreliable.

A chatbot is therefore an excellent probe and a poor certificate. Move from talk to work. Give the system novel tasks, tools, feedback, long horizons and opportunities to discover that its first interpretation was wrong. Generality appears in adaptation and completed outcomes, not in how human the first paragraph feels.

“Human-level is one finish line”

The phrase suggests an average person represented by one score. No such person exists. Human ability varies by domain, training, age, language, culture and the tools available. A competent nurse, solicitor, mechanic and mathematician carry different profiles. Even one person can be expert at work and hopeless at navigation.

Machines widen the mismatch. A system may exceed specialists in a formal task and fall below a child in physical adaptation. Combining those results

into one level produces a number with no stable meaning. It also hides whether the comparison concerns accuracy, speed, cost, learning or independence.

The ladder persists because it gives forecasts a clean milestone. Investors, journalists and policymakers can attach a date to human-level AI more easily than to a multidimensional profile. The simplicity is purchased by deleting the information needed for action.

Replace the line with a map. State the domains, reference population, conditions, tools, reliability and time horizon. Separate breadth from depth and capability from autonomy. A system can have economically transformative coverage before matching human competence everywhere, while one superhuman peak can remain narrow. The map may deny us a dramatic crossing. It tells us where the machine can be trusted.

“One benchmark will settle it”

A benchmark offers what the argument lacks: a score, a leaderboard and a visible winner. The temptation is to crown the first system that reaches a human baseline. That works only if the test samples the property cleanly, remains independent of training and predicts performance beyond itself.

AGI makes each condition difficult. Public tasks become optimisation targets. Training data may contain exact items or close relatives. A specialised scaffold can exploit the format. Human baselines depend on instructions and tools. A system can master the test family while failing in open environments with missing goals and delayed feedback.

This does not make benchmarks useless. Controlled tests isolate mechanisms, expose progress and permit comparison. ARC-style tasks, coding suites and professional examinations each reveal something. They reveal different things. A provider-specific interface may measure the strongest product; a common interface may compare models more fairly.

The correction is convergence. Demand fresh and concealed tasks, varied domains, learning curves, long projects, calibration, adversarial attack and

deployment evidence. Publish costs, assistance and failures. The stronger the claim, the less it should depend on one number. A genuine threshold will leave traces across many tests. It should not disappear when the examiner changes the paper.

“More scale guarantees AGI”

Scaling has earned respect. Across important ranges, more suitable combinations of data, model capacity and computation have improved performance predictably. Broad pretraining has produced systems able to reuse one foundation across many tasks. Dismissing that record because the mechanism feels brute-force is poor analysis.

A trend is not a guarantee. Scaling laws are empirical relations under stated architectures, data and budgets. They do not prove that continual learning, robust planning, causal understanding, calibration or physical competence will arrive before cost, data quality or engineering bottlenecks bind. Some reported sudden abilities also become smoother when measured with less brittle metrics.

The opposite claim, that scale can never yield generality because it lacks a favoured ingredient, is equally ahead of the evidence. Large learned systems may acquire internal methods that were not designed one by one, and tools can compensate for weaknesses. Hybrid and world-model approaches may also use scale.

The useful forecast names the bottleneck. Which ability is missing? Why should more training remove it, or why should it persist? What result would change the view? “It gets better when larger” is evidence for continued gains. “Therefore every required property appears” is faith with a graph attached.

“AGI will arrive in one obvious moment”

The image is a laboratory countdown followed by a machine waking up. Real technological thresholds are often administrative before they are dramatic. Capabilities spread across versions, tools and organisations. One system may cross a cyber threshold, another a scientific one, while neither can manage an ordinary long project without rescue.

Definitions also move. Conversation once looked decisive. Then broad question answering, professional examinations and agentic work became candidate lines. Each success reveals a harder remainder. This can be dishonest goalpost shifting, but it can also be legitimate learning about what the earlier test failed to capture.

Commercial effects make the arrival more gradual. Firms automate tasks, redesign jobs and grant systems access without waiting for consensus. A system may become economically general across large parts of digital work before philosophers accept the label. Conversely, a laboratory may announce AGI while customers still need extensive supervision.

There may eventually be a release whose breadth and reliability force rapid agreement. We cannot assume it. Track capability profiles and deployment envelopes instead of waiting for a ceremony. The date historians choose later may mark a paper, product, benchmark, market shift or governance decision. People living through the change may disagree for years.

“The only futures are utopia or extinction”

AGI debate is pulled towards two clean endings. In one, abundant intelligence cures disease, raises productivity and removes drudgery. In the other, an uncontrollable system ends human control or human life. Both identify real possibilities. Neither describes the full decision space between them.

More ordinary outcomes can still be enormous. Broad systems may improve research while concentrating ownership. They may make expertise cheaper while weakening professions, expand education while enlarging

surveillance, strengthen defenders while lowering the cost of attack, and raise output while distributing gains badly. Institutions can adapt unevenly, producing different futures across countries and sectors.

Extreme risk deserves analysis because low probability does not cancel catastrophic consequence. Optimism deserves evidence because scientific and material benefits could be substantial. The error is using either endpoint to avoid studying mechanisms, probabilities and interventions.

Ask what capability appears, who controls it, what access it receives, how failure propagates and which safeguards remain effective under competition. Many choices are available before a system is built, during deployment and after incidents. The future is not a vote between salvation and doom. It is a sequence of technical and political allocations, each changing the next set of options.

Use It

Test the change, not the familiar task

When someone claims general intelligence, ask what changed between training and evaluation. A new question drawn from the same template tests competence inside a family. A changed goal, interface, rule, language or environment tests whether the method transfers.

This does not require inventing an impossible surprise. Give the system a task whose surface is unfamiliar but whose underlying demands a competent person could infer. Limit demonstrations, then measure how quickly performance improves. Change one assumption after progress begins. See whether the system notices, clings to the old model or merely produces a new confident answer.

The same lens works in procurement. Do not begin with the supplier's showcase. Select cases from your future workflow, especially the awkward boundary cases your current process struggles to classify. Hold some back until the system design is settled. After a successful pilot, change volume,

wording, user group or source quality. Generality is valuable because the world refuses to preserve the pilot.

Separate the model, system, agent and institution

Draw four boxes before judging any claim. The model produces judgements. The system supplies prompts, retrieval, memory, tools and filters. The agent carries work across decisions. The institution grants access, sets incentives, reviews output and carries liability.

Then ask which box produced the gain. A model update may matter less than a better search index. A memory layer may rescue long tasks while introducing stale facts. Human reviewers may be doing the difficult cases invisibly. Standardised digital interfaces may be supplying structure that disappears in another language, institution or physical setting. A striking benchmark may depend on a provider adapter unavailable elsewhere.

This accounting prevents two errors. You will not reject a useful product because the base model is imperfect if the whole system catches the relevant failures. You will not credit the model for performance that depends on expensive human rescue. More importantly, you can change the correct layer. Sometimes the answer is better training. Sometimes it is narrower permission, better data, a formal checker or a person with authority to stop the run.

Measure the whole chain, not the best answer

AGI claims attract spectacular examples because one success is easy to show. Real value depends on the chain from instruction to consequence. Measure whether the system understood the goal, found the right evidence, used tools correctly, detected failure, produced a usable result and left a record another person can audit.

Use distributions rather than selected outputs. Repeat runs. Record time, cost, retries and human corrections. Separate harmless defects from failures that reverse the decision. One wrong date in a draft and one fabricated authority in legal advice are both errors, but they do not belong in

the same average.

Long tasks deserve checkpoints based on external state. Did the file change as intended? Did the calculation reconcile? Did the customer receive the right item? A verbal claim that the work is finished is not proof. The stronger and more autonomous the system becomes, the more evaluation should move from judging prose to inspecting the world after the action.

Track permissions, reversibility and blast radius

Do not wait for a verdict on AGI before controlling authority. List what the system can read, disclose, modify, purchase, publish and delete. Then mark which actions are reversible, which can be reviewed before execution and how many people or assets one error can reach.

Increase autonomy one rung at a time. Suggestion can become drafting, drafting can become queued action, and queued action can become execution under limits. Each rung should earn evidence from the previous one. A system that writes excellent emails has not thereby earned permission to send them to every customer.

Reversibility changes tolerance. A wrong internal summary can be corrected. A transferred payment, leaked secret or altered production system may not be recoverable. Keep credentials scoped, separate environments, set spending and time limits, preserve logs and design a route back to a known good state. These controls are useful whether the model is narrow, general or something people are still arguing about.

Use capability triggers, not calendar forecasts

Plans built around “AGI in 2030” confuse a forecast with an operating condition. Timelines vary because definitions, methods and incentives vary. Replace the date with observable triggers tied to your exposure.

A research organisation might watch for systems that can reproduce a week's experimental analysis with low supervision. A software firm might track the fraction of repository-level tasks completed with tests and review. A regulator may care when models cross defined cyber, biological or

autonomy thresholds. A school may care when unsupervised systems can complete its assessments while masking the assistance.

For each trigger, decide in advance what changes: access, evaluation, staffing, insurance, incident reporting or independent review. Update the trigger as evidence improves. This approach remains useful if progress is slower, faster or stranger than predicted. It also reduces the temptation to reinterpret every release as proof that one's favourite date is approaching.

Preserve option value and human competence

Under uncertainty, avoid designs that make one forecast irreversible. Do not rebuild a critical operation so completely around one provider that failure or withdrawal becomes unmanageable. Keep data portable, interfaces documented and a degraded mode that people can run. Test whether the human fallback still works rather than assuming it survives through disuse.

Human competence can decay when the machine handles routine cases and leaves only rare difficult ones. Reviewers then receive less practice precisely where judgement matters most. Rotate people through unassisted work, preserve training cases and make escalation a real activity rather than a button nobody uses. Automation should remove needless labour without deleting the ability to recognise when it has gone wrong.

Option value also applies to policy. Rules can tighten access as demonstrated capability rises without pretending that every model poses the same risk. Independent evaluation capacity, secure research environments and incident-sharing arrangements remain useful across several futures. Building the ability to observe and respond is safer than betting everything on one predicted architecture.

The limits

No framework in this book can turn an open research question into a settled timetable. Evaluation samples behaviour. It cannot prove that every future task, prompt, tool or adversary will produce the same result. Secret training data and proprietary systems restrict independent scrutiny. Providers and critics both select examples, and rapid updates can make a careful judgement stale.

The model also stops before several neighbouring subjects. It does not decide whether machines can be conscious, provide a full theory of human intelligence, derive alignment mathematically or forecast wages and growth. It cannot tell one organisation which legal duties apply or which deployment is acceptable. Those decisions depend on current law, sector, population and consequence.

Most of all, a threshold does not determine a policy. Two people can agree on a system's capabilities and disagree about ownership, distribution, access and acceptable risk. Technical evidence narrows the argument. It does not remove politics or values from it.

The one thing to keep

Keep the changed problem.

A narrow machine can look universal while the world stays arranged around its strengths. A broad model can look foolish when one assumption moves. Neither impression settles the case. The useful evidence is what remains after the task stops resembling the preparation.

Ask what the system had to acquire rather than recall. Ask whether it built a model, tested it, noticed failure and transferred the lesson. Ask whether success survived a different interface, language, time horizon or objective. Then name the tools and people that made the result possible. General competence belongs to the arrangement that keeps adapting, not to whichever component receives the applause.

Carry the same test into deployment. Change of problem is not confined to

benchmarks. Customers behave differently once a system makes decisions. Attackers study its rules. Workers reorganise around it. Generated material enters future datasets. The machine alters the environment from which its next evidence arrives. A one-time pass therefore cannot license unlimited authority.

AGI may eventually become obvious through accumulated performance. It may arrive as a disputed label attached to a gradual economic transformation. It may remain the name of a horizon while systems with no agreed status acquire immense power. In every case, the discipline is the same.

Do not ask first whether the machine is generally intelligent. Ask what evidence would survive a change of problem, and what the system is allowed to do while the answer is still uncertain.

Terms

Artificial general intelligence. A disputed name for AI with broad, dependable competence across many cognitive tasks, especially the ability to acquire and transfer skills beyond cases prepared in advance.

Narrow AI. A system designed or effective within a bounded task or domain. Superhuman performance can remain narrow when rules, inputs and objectives stay fixed.

General-purpose AI. A model or system usable across a wide variety of tasks. The term describes breadth of application without claiming human-level reliability or full AGI.

Human-level AI. AI compared with a stated human group across stated tasks and conditions. Without the population, tools, cost and performance measure, the phrase is incomplete.

Artificial superintelligence. A hypothetical system that substantially exceeds human cognitive performance across almost all relevant domains. It is conceptually beyond AGI, though definitions and routes remain

contested.

Capability. What a system can accomplish under specified conditions.

Tools, prompting, time, cost and elicitation matter, so capability is measured rather than inferred from a model name. Failed elicitation can understate it; favourable scaffolding can overstate portability.

Generality. Breadth joined to transfer and adaptation. A general system applies useful knowledge across domains and makes progress when the task, goal or environment changes.

Performance. The depth of competence shown on a task distribution, including accuracy, reliability, speed and cost. High breadth with poor performance does not establish useful AGI.

Transfer. Improvement on a new task because knowledge learned elsewhere remains relevant. Transfer distinguishes reusable structure from specialisation to one format or dataset.

Out-of-distribution. Describes cases drawn from conditions different from those represented in training or routine testing. Performance may fall when language, users, rules or environments shift.

Continual learning. Adding knowledge or skills over time while preserving earlier competence. It matters because a general system cannot require complete retraining after every change.

Catastrophic forgetting. Severe loss of earlier abilities when a model learns new material. The problem complicates continual learning and shows that adaptation can carry hidden costs.

Metacognition. Control over one's own problem-solving process, including recognising uncertainty, selecting strategies and deciding when to ask, verify, persist or stop. Functional self-knowledge improves reliability.

Calibration. The match between expressed confidence and observed correctness. A calibrated system is uncertain more often when it is wrong, making review and delegation safer. Confidence wording alone is not

evidence of calibration.

World model. An internal representation that predicts how an environment changes, especially after actions. A useful model represents uncertainty as well as expected change. It supports planning, imagined trials and correction when observations disagree.

Foundation model. A broadly trained model adapted or connected to many downstream uses. One foundation can spread capability, defects and dependencies across numerous products. This shared base creates efficiency and correlated failure.

Agent. A system that pursues an objective through a sequence of observations and actions. Agents may plan, use tools, preserve memory, revise and operate with varying independence. The objective and stopping rule still come from somewhere.

Autonomy. How far a system can continue by choosing and carrying out intermediate actions before another human instruction. Limits may apply to duration, action type, spend or approval. Greater autonomy enlarges the length and branching of the action chain.

Authority. Permission and effective power to affect people, data, money or physical systems. Authority comes from access and human deference, not from intelligence alone. A weak system with wide credentials can still cause large harm.

Tool use. Calling calculators, search, code, databases, software or physical devices to obtain evidence and act. Tools can extend competence while adding security and control risks. Results should be checked against external state rather than accepted as narrated.

Scaffold. External structure around a model, such as prompts, memory, planners, retries or specialised checkers. Scaffolds can change results enough that evaluations must disclose them. The strongest product may depend on a scaffold unavailable to competitors.

Benchmark. A standard task set and scoring method used to compare

systems. Benchmarks isolate abilities but weaken when contaminated, optimised against or mistaken for real work. A suite of benchmarks provides stronger evidence than one score.

Contamination. Overlap between evaluation material and training or development data. Exact leakage is clearest, but repeated tuning to a public task family can also erode independence. Private hold-outs reduce but do not eliminate the problem.

Ecological validity. The degree to which a test reflects the conditions that matter outside it. Clean tasks aid comparison while omitting ambiguity, interruptions, incentives and consequences. Deployment evidence can expose failures that controlled tests miss.

Goodhart's law. The warning that a measure can stop tracking its intended goal when it becomes a target. Optimisation finds shortcuts, loopholes and proxy failures.

Alignment. The problem of keeping AI behaviour within intended goals, constraints and values. It includes training, system design, permissions, monitoring and the institutions choosing the intention.

Misalignment. Behaviour that departs materially from intended objectives or boundaries. It can arise from poor specification, generalisation, incentives, deception, error or conflict among human goals.

Corrigibility. A system's willingness and ability to accept correction, interruption or modification without resisting the change. The concept matters most when agents pursue long objectives.

Interpretability. Methods for understanding internal representations or the causes of behaviour. Partial insight can support diagnosis, but no current method makes every important decision transparent.

Takeoff. A period in which AI capability improves rapidly, potentially because AI accelerates AI research. Speed depends on which software, physical and institutional bottlenecks remain.

Go Deeper

The accessible overview

Melanie Mitchell, *Artificial Intelligence: A Guide for Thinking Humans* (2019). Mitchell explains the field through its recurring ambitions, working systems and missing common sense. The book predates the latest agentic models, so its product examples are no longer current. Its central discipline remains useful: separate a machine's observed competence from the human understanding we project onto it. Read it for the historical and conceptual foundation this book has compressed. Its treatment of analogy, abstraction and the "barrier of meaning" supplies a strong sceptical counterweight to accounts that infer understanding from fluent output alone.

The primary source

Alan Turing, "Computing Machinery and Intelligence" (1950). The famous imitation game occupies only part of a much richer paper on learning machines, objections, evidence and the danger of defining thought by decree. It is readable without mathematics and short enough to encounter directly. Read it to see how an operational test can clarify a confused question, and how even an excellent test can later be overtaken by the systems it helped inspire. Notice that Turing predicts education and child machines rather than insisting on a finished adult mind built from explicit rules. That part of the paper has aged better than the popular shorthand.

The classification framework

Meredith Ringel Morris and colleagues, “Position: Levels of AGI for Operationalizing Progress on the Path to AGI” (2024). This position paper separates performance, generality and autonomy, then proposes levels rather than one binary threshold. The exact matrix is contestable, as any AGI taxonomy must be, but its questions are disciplined and practical. Read it when a claim such as human-level or general needs to be converted into dimensions, reference groups and evidence. It also keeps deployment autonomy separate from raw capability, a distinction that becomes indispensable when the same model can be offered as a passive assistant or a long-running agent.

The current evidence map

International AI Safety Report 2026, led by Yoshua Bengio and developed with more than one hundred contributing experts. It surveys current general-purpose capabilities, evaluation gaps, misuse, control risks and risk-management methods while marking disagreement and uncertainty. It is long and institutional rather than narrative. Read the executive summary first, then the capability and evaluation sections. Use it as a dated map, not a prophecy. Its references provide a route into the underlying studies when one finding matters. The report is especially valuable when confident claims draw from one laboratory result, because it shows where experts converge, where evidence is thin and why deployment can outrun evaluation.

Notes and Sources

Verification date and status

This manuscript was source-checked on 4 September 2026. Artificial general intelligence has no single widely accepted definition or certification test. Claims about whether any present system qualifies are therefore attributed to the definition, evaluator and operating conditions involved. Current model results are snapshots of named versions and may change through updates, access, prompting, tools and further independent testing. Publication date, model version, evaluation period and data vintage are kept separate: the *International AI Safety Report 2026* synthesises evidence published before December 2025, while the Astra system card and ARC Prize results were published on 3 September 2026. The six-desk opening scenario is hypothetical, not a report of an unnamed deployment. Other apparently factual cases are documented; mechanism examples are written as conditional or illustrative.

The current threshold example

The opening case uses OpenAI's *GPT-6 Astra System Card*, published on 3 September 2026. The provider states that Astra is its first broadly deployed model to reach the Critical level for cybersecurity capability under its Preparedness Framework, while not reaching the framework's High threshold for AI self-improvement. The same card reports stronger results on several alignment evaluations, lower chain-of-thought and full-context monitorability across most tested lengths, and higher action-only monitorability in aggregate. It also explains that some action-only gains reflected easier-to-detect violations or comparison artefacts rather than clearer reasoning. These are provider assessments with disclosed limitations, not an independent declaration of AGI or proof that every deployment behaves the same way.

Ina Fried's Axios report of the product briefing on 3 September 2026 says that OpenAI president Greg Brockman personally believed the company had reached AGI, described Astra as a generational leap, and said it might

later be seen as AGI's arrival. The manuscript retains this as an attributed executive judgement because it establishes that the label was attached to a live release. It is not used as evidence that a technical threshold was crossed.

ARC-AGI-3 and the two interfaces

The figures of 62.7 per cent and 99.9 per cent come from Greg Kamradt's ARC Prize report, "OpenAI's GPT-6 Astra on ARC-AGI-3", published on 3 September 2026. The 62.7 per cent result was Astra at maximum reasoning effort on the Semi-Private set using ARC Prize's standard, provider-neutral interface. The 99.9 per cent result was Astra at high reasoning effort using a provider adapter that preserved opaque reasoning state between requests and used context compaction. The corresponding reported costs were about \$26,098 and \$18,817. Costs are not quoted in the body because they are unusually sensitive to pricing, implementation and exchange context.

ARC Prize reports that Astra using the Provider Adapter at maximum reasoning effort used fewer actions than its median tested human baseline on 96 per cent of completed levels. The body retains the exact condition and percentage. The human comparison concerns action efficiency within the benchmark, not total energy, monetary cost, life experience or open-world competence.

ARC Prize defines AGI through human-efficient skill acquisition, but explicitly says that saturating ARC-AGI-3 would not prove AGI. Its environments are novel, abstract and interactive, yet tightly bounded, deterministic and closed-ended. The standard interface and provider adapter answer different questions, and some tool-rich machine conditions were not equivalent to the human testing conditions. These qualifications are central to the argument rather than afterthoughts.

Definitions and dimensions

OpenAI's Charter defines AGI as highly autonomous systems that outperform humans at most economically valuable work. Meredith Ringel Morris and colleagues instead propose levels based on performance and generality, with autonomy considered separately for deployment. François Chollet treats intelligence as skill-acquisition efficiency and created the Abstraction and Reasoning Corpus to test broad generalisation from few examples. Shane Legg and Marcus Hutter formalise intelligence as an agent's ability to achieve goals across a wide range of environments. The coexistence of these definitions supports the book's claim that different operational choices can produce different threshold dates.

The statement that the name gained wider currency in the 2000s is supported by Ben Goertzel and Cassio Pennachin's 2007 edited volume, *Artificial General Intelligence*. It is not presented as a claim that they were the first people to use the phrase. The 2023 “sparks” episode follows Sébastien Bubeck and colleagues, who argued that an early GPT-4 version displayed broad but incomplete general intelligence. The manuscript reports the claim and the methodological dispute, not its conclusion.

The distinction between breadth and depth follows Morris and colleagues. The emphasis on acquiring new skill rather than counting stored skills follows Chollet. The book's central formula, that generality appears when the problem changes, is an editorial synthesis of transfer, out-of-distribution performance, adaptation and skill-acquisition efficiency. It is a practical test, not a proposed universal mathematical definition.

Human comparison and jagged capability

The *International AI Safety Report 2026* describes current general-purpose AI capability as jagged: leading systems can excel on difficult tasks while failing on apparently simpler ones, longer workflows, basic physical tasks or languages other than English. The report also stresses that test performance often does not predict real-world utility or risk, calling this an evaluation gap. Its analysis draws on scientific, technical and socioeconomic evidence published before December 2025, so the manuscript combines that dated synthesis with sources published on 3 September 2026 for the Astra case.

The claim that human-level intelligence is a profile rather than one line does not deny the evidence for general cognitive factors in human psychometrics. It says that an AGI declaration still needs a task distribution, population, conditions and measure. Human cognition, expertise and social responsibility are covered only as comparison problems here; a full account belongs to *Intelligence in a Hurry* and related books.

From symbolic systems to broad learned models

Turing's 1950 paper supplies the imitation game, the learning-machine proposal and the operational turn away from defining thought directly. The 1955 Dartmouth proposal supplies the founding ambition that aspects of intelligence could be described precisely enough for machine simulation. Allen Newell, J. C. Shaw and Herbert Simon's General Problem Solver supplies means-ends analysis. Douglas Lenat and R. V. Guha supply the Cyc example of an attempt to encode commonsense knowledge at scale.

The ImageNet example follows Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton's 2012 paper. Deep Blue's 1997 victory is used only to illustrate superhuman narrow competence. AlphaGo's combination of learned policy and value networks with tree search follows David Silver and colleagues' 2016 *Nature* paper. None of these milestones is presented as a straight, inevitable road to AGI.

Scaling, emergence and alternative roads

The summary of neural-language-model scaling follows Jared Kaplan and colleagues. The point that model and data scale must be balanced under a compute budget follows Jordan Hoffmann and colleagues' Chinchilla work. Rylan Schaeffer, Brando Miranda and Sanmi Koyejo show how thresholded metrics can make smooth performance changes look discontinuous. These papers support observed regularities and measurement cautions. They do not establish that scaling must either achieve or fail to achieve AGI.

The world-model route draws on David Ha and Jürgen Schmidhuber, Yann LeCun's proposed architecture for autonomous machine intelligence, and Joshua Tenenbaum and colleagues' broader programme for machines that learn more like people. Rodney Brooks supplies the historical challenge that intelligence may depend on situated interaction rather than detached symbolic representation. These are rival research programmes, not settled prerequisites.

The continual-learning discussion follows German Parisi and colleagues' review of lifelong learning and catastrophic forgetting. External memory can preserve task state without changing model parameters, but that is not treated as equivalent to integrating a new skill permanently.

Agents, memory and long-horizon work

ReAct, by Shunyu Yao and colleagues, interleaves reasoning and action. Joon Sung Park and colleagues' generative agents use stored experiences, retrieval, reflection and planning. Guanzhi Wang and colleagues' Voyager develops a growing library of executable skills in Minecraft. These works are cited as bounded demonstrations of agent components, not evidence of open-world AGI. Thomas Kwa and colleagues' software-task study supplies the time-horizon example. Its task mix and authors' explicit external-validity warning prevent the result being generalised to ordinary work.

The illustration that 99 per cent step reliability across one hundred independent, fatal steps gives roughly a 37 per cent clean-chain probability is the calculation 0.99 raised to the power 100. Real failures are not

necessarily independent, steps differ, and capable systems can detect or repair some mistakes. The figure explains compounding rather than predicting the reliability of a named product.

Benchmarks and the target problem

Goodhart's law is used in the broad policy sense associated with Charles Goodhart and later popularised in an audit context by Marilyn Strathern: pressure on a measure can separate it from the underlying objective. Benchmark contamination is only one failure mode. Repeated development against a public task family can also weaken independence without exact item leakage.

The claim that evaluation should combine controlled tests, held-back tasks, adversarial attempts, long-horizon work and deployment evidence is a synthesis. It is consistent with the International AI Safety Report's evaluation-gap analysis, Morris and colleagues' benchmark discussion, ARC Prize's evolving test series, and modern evaluation programmes such as HELM. No listed source claims that one fixed portfolio can certify all future behaviour.

Alignment, authority and control

Dario Amodei and colleagues' "Concrete Problems in AI Safety" supplies examples of specification, robustness and monitoring failures in learning systems. Lauro Langosco Di Langosco and colleagues' goal-misgeneralisation work distinguishes learning the training signal from pursuing the intended goal in new settings. Alexander Turner and colleagues analyse why reward-optimal policies can have incentives related to power in formal environments. These results motivate caution; they do not show that every capable system will seek power or deceive.

The separation of capability, autonomy and authority is partly adapted from Morris and colleagues and partly an editorial operating model. Authority includes credentials, permissions, scale and human deference. The layered-control recommendations are supported by the 2026 safety report's defence-in-depth discussion and NIST's AI Risk Management Framework.

They are general risk-management principles, not legal advice for any jurisdiction or sector.

Consciousness

The distinction between general capability and consciousness follows the separation between functional performance and subjective experience. Patrick Butlin and colleagues review several scientific theories of consciousness and derive indicators for AI assessment. They do not offer a conclusive consciousness test and do not claim current systems satisfy a complete case. The manuscript therefore treats machine consciousness as open and neighbouring, not as a requirement for AGI.

Self-improvement, takeoff and consequence

I. J. Good's 1965 essay supplies the classic intelligence-explosion argument. The manuscript weakens the popular shorthand by separating software improvement from hardware, fabrication, energy, experiments, security, capital and institutional decision-making. The claim that AI could accelerate AI research is retained as conditional. Astra's provider assessment below the High AI self-improvement threshold is one current data point, not a ceiling on later systems.

Claims about science, work, security, concentration and verification are mechanism-level possibilities rather than point forecasts. The *International AI Safety Report 2026* supports the existence of expanding scientific and cyber capabilities, uneven adoption, information asymmetries and limited evidence on many safeguards. The book does not estimate an AGI arrival date, employment total, probability of catastrophe or rate of economic growth.

Bibliography

Primary and original sources

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman and Dan Mané. “Concrete Problems in AI Safety.” arXiv:1606.06565, 2016.

ARC Prize Foundation. “ARC-AGI-3: A New Challenge for Frontier Agentic Intelligence.” arXiv:2603.24621, version 2, 2026.

Brooks, Rodney A. “Intelligence Without Representation.” *Artificial Intelligence* 47, 1991.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro and Yi Zhang. “Sparks of Artificial General Intelligence: Early Experiments with GPT-4.” arXiv:2303.12712, 2023.

Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon and Rufin VanRullen. “Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.” arXiv:2308.08708, version 3, 2023.

Campbell, Murray, A. Joseph Hoane Jr and Feng-hsiung Hsu. “Deep Blue.” *Artificial Intelligence* 134, 2002.

Chollet, François. “On the Measure of Intelligence.” arXiv:1911.01547, 2019.

Goertzel, Ben and Cassio Pennachin, editors. *Artificial General Intelligence*. Berlin and Heidelberg: Springer, 2007.

Good, I. J. “Speculations Concerning the First Ultraintelligent Machine.” *Advances in Computers* 6 (1965): 31-88.

Goodhart, Charles A. E. “Problems of Monetary Management: The UK Experience.” *Papers in Monetary Economics* 1. Sydney: Reserve Bank of Australia, 1975.

Ha, David and Jürgen Schmidhuber. “World Models.” arXiv:1803.10122, 2018.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas and others. “Training Compute-Optimal Large Language Models.” arXiv:2203.15556, 2022.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu and Dario Amodei. “Scaling Laws for Neural Language Models.” arXiv:2001.08361, 2020.

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems* 25, 2012.

Kwa, Thomas, Ben West, Joel Becker, Amy Deng and others. “Measuring AI Ability to Complete Long Software Tasks.” *NeurIPS 2025*; arXiv:2503.14499, version 4, 10 July 2026.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum and Samuel J. Gershman. “Building Machines That Learn and Think Like People.” *Behavioral and Brain Sciences* 40, 2017.

Langosco Di Langosco, Lauro, Jack Koch, Lee D. Sharkey, Jacob Pfau and David Krueger. “Goal Misgeneralization in Deep Reinforcement Learning.” *Proceedings of the 39th International Conference on Machine Learning*, 2022.

LeCun, Yann. “A Path Towards Autonomous Machine Intelligence.” OpenReview, 2022.

Legg, Shane and Marcus Hutter. “Universal Intelligence: A Definition of

Machine Intelligence.” *Minds and Machines* 17, 2007.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.” 1955.

Morris, Meredith Ringel, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet and Shane Legg. “Position: Levels of AGI for Operationalizing Progress on the Path to AGI.” *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:36308-36321, 2024.

Newell, Allen, J. C. Shaw and Herbert A. Simon. “Report on a General Problem-Solving Program.” *Proceedings of the International Conference on Information Processing*, 1959.

OpenAI. *OpenAI Charter*. 2018.

OpenAI. *GPT-6 Astra System Card*. 3 September 2026.

Parisi, German I., Ronald Kemker, Jose L. Part, Christopher Kanan and Stefan Wermter. “Continual Lifelong Learning with Neural Networks: A Review.” *Neural Networks* 113, 2019.

Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang and Michael S. Bernstein. “Generative Agents: Interactive Simulacra of Human Behavior.” *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

Schaeffer, Rylan, Brando Miranda and Sanmi Koyejo. “Are Emergent Abilities of Large Language Models a Mirage?” *Advances in Neural Information Processing Systems* 36, 2023.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche and others. “Mastering the Game of Go with Deep Neural Networks and Tree Search.” *Nature* 529, 2016.

Turing, A. M. “Computing Machinery and Intelligence.” *Mind* 59, no. 236, 1950.

Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch and Prasad Tadepalli. “Optimal Policies Tend to Seek Power.” *Advances in Neural Information Processing Systems* 34, 2021.

Wang, Guanzhi, Yuqi Xie, Yunfan Jiang, Ajay Mandekar, Chaowei Xiao, Yuke Zhu, Linxi Fan and Anima Anandkumar. “Voyager: An Open-Ended Embodied Agent with Large Language Models.” arXiv:2305.16291, 2023.

Yao, Shunyu, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan and Yuan Cao. “ReAct: Synergizing Reasoning and Acting in Language Models.” *International Conference on Learning Representations*, 2023.

Current syntheses, evaluations and further reading

Bengio, Yoshua, Stephen Clare, Carina Prunkl, Malcolm Murray and others. *International AI Safety Report 2026*. DSIT 2026/001. London: Department for Science, Innovation and Technology, 2026.

Fried, Ina. “‘Welcome to the AGI Era,’ OpenAI Says as GPT-6 Astra Debuts.” *Axios*, 3 September 2026.

Kamradt, Greg. “OpenAI’s GPT-6 Astra on ARC-AGI-3.” ARC Prize, 3 September 2026.

Lenat, Douglas B. and R. V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Reading, MA: Addison-Wesley, 1990.

Liang, Percy, Rishi Bommasani, Tony Lee and others. “Holistic Evaluation of Language Models.” arXiv:2211.09110, 2022.

Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux, 2019.

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: NIST, 2023.

Strathern, Marilyn. "'Improving Ratings': Audit in the British University System." *European Review* 5, no. 3, 1997.