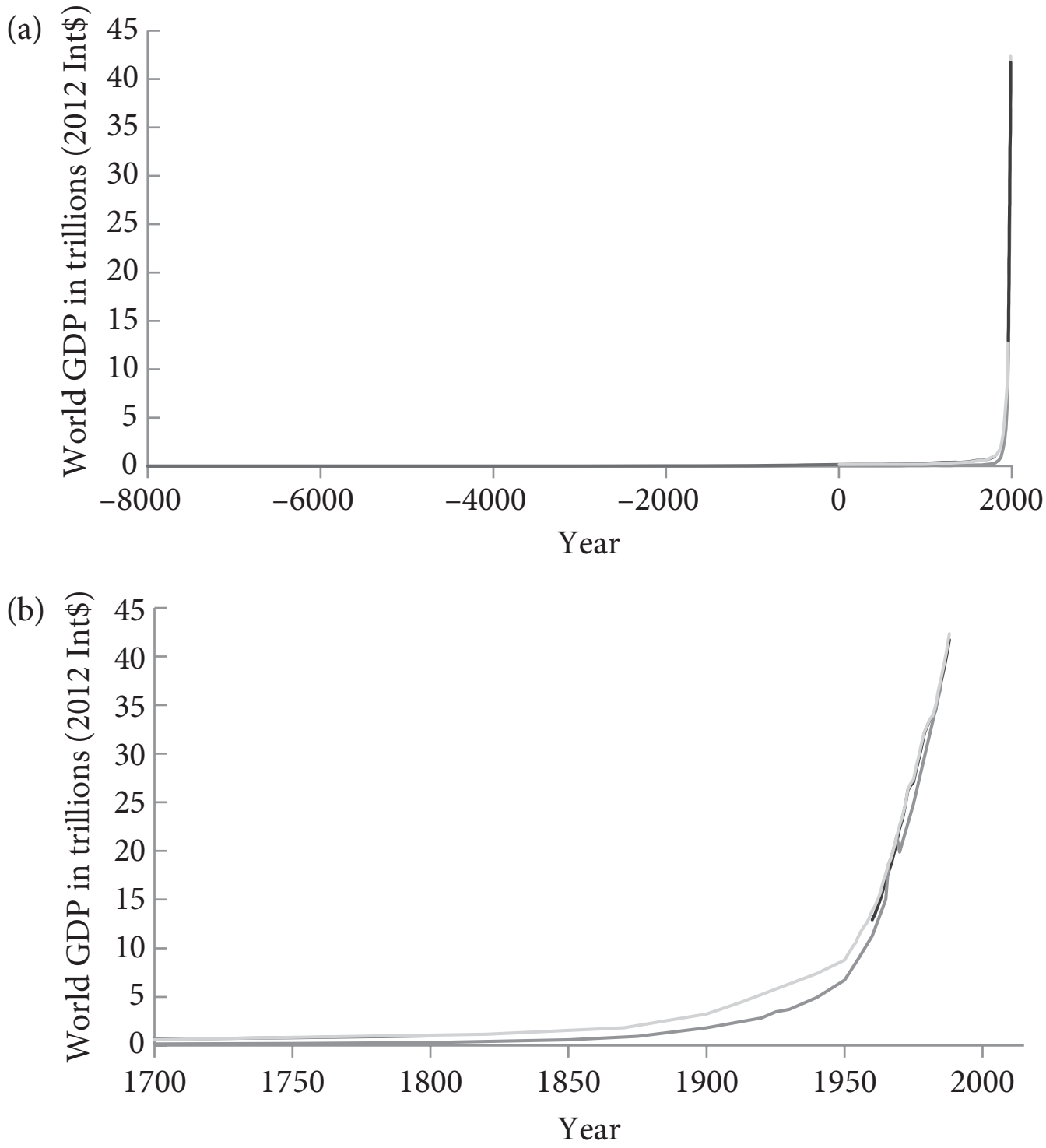

Figure 1 Long-term history of world GDP. Plotted on a linear scale, the history of the world economy looks like a flat line hugging the *x*-axis, until it suddenly spikes vertically upward. (a) Even when we zoom in on the most recent 10,000 years, the pattern remains essentially one of a single 90° angle. (b) Only within the past 100 years or so does the curve lift perceptibly above the zero-level. (The different lines in the plot correspond to different data sets, which yield slightly different estimates.[6])

Table 1 ***Game-playing AI***

Checkers	Superhuman	Arthur Samuel's checkers program, originally written in 1952 and later improved (the 1955 version incorporating machine learning), becomes the first program to learn to play a game better than its creator.[37] In 1994, the program CHINOOK beats the reigning human champion, marking the first time a program wins an official world championship in a game of skill. In 2002, Jonathan Schaeffer and his team "solve" checkers, i.e. produce a program that always makes the best possible move (combining alpha-beta search with a database of 39 trillion endgame positions). Perfect play by both sides leads to a draw.[38]
Backgammon	Superhuman	1979: The backgammon program BKG by Hans Berliner defeats the world champion—the first computer program to defeat (in an exhibition match) a world champion in any game—though Berliner later attributes the win to luck with the dice rolls.[39] 1992: The backgammon program TD-Gammon by Gerry Tesauro reaches championship-level ability, using temporal difference learning (a form of reinforcement learning) and repeated plays against itself to improve.[40] In the years since, backgammon programs have far surpassed the best human players.[41]
Traveller TCS	Superhuman in collaboration with human[42]	In both 1981 and 1982, Douglas Lenat's program Eurisko wins the US championship in Traveller TCS (a futuristic naval war game), prompting rule changes to block its unorthodox strategies.[43] Eurisko had heuristics for designing its fleet, and it also had heuristics for modifying its heuristics.
Othello	Superhuman	1997: The program Logistello wins every game in a six-game match against world champion Takeshi Murakami.[44]
Chess	Superhuman	1997: Deep Blue beats the world chess champion, Garry Kasparov. Kasparov claims to have seen glimpses of true intelligence and creativity in some of the computer's moves.[45] Since then, chess engines have continued to improve.[46]
Crosswords	Expert level	1999: The crossword-solving program Proverb outperforms the average crossword-solver.[47]

Table 1 *Continued*

		2012: The program Dr. Fill, created by Matt Ginsberg, scores in the top quartile among the otherwise human contestants in the American Crossword Puzzle Tournament. (Dr. Fill's performance is uneven. It completes perfectly the puzzle rated most difficult by humans, yet is stumped by a couple of nonstandard puzzles that involved spelling backwards or writing answers diagonally.)[48]
Scrabble	Superhuman	As of 2002, Scrabble-playing software surpasses the best human players.[49]
Bridge	Equal to the best	By 2005, contract bridge playing software reaches parity with the best human bridge players.[50]
Jeopardy!	Superhuman	2010: IBM's *Watson* defeats the two all-time-greatest human *Jeopardy!* champions, Ken Jennings and Brad Rutter.[51] *Jeopardy!* is a televised game show with trivia questions about history, literature, sports, geography, pop culture, science, and other topics. Questions are presented in the form of clues, and often involve wordplay.
Poker	Varied	Computer poker players remain slightly below the best humans for full-ring Texas hold 'em but perform at a superhuman level in some poker variants.[52]
FreeCell	Superhuman	Heuristics evolved using genetic algorithms produce a solver for the solitaire game FreeCell (which in its generalized form is NP-complete) that is able to beat high-ranking human players.[53]
Go	Very strong amateur level	As of 2012, the Zen series of go-playing programs has reached rank 6 dan in fast games (the level of a very strong amateur player), using Monte Carlo tree search and machine learning techniques.[54] Go-playing programs have been improving at a rate of about 1 dan/year in recent years. If this rate of improvement continues, they might beat the human world champion in about a decade.

Table 2 ***When will human-level machine intelligence be attained?***[81]

	10%	50%	90%
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Combined	2022	2040	2075

Table 3 ***How long from human level to superintelligence?***

	Within 2 years after HLMI	Within 30 years after HLMI
TOP100	5%	50%
Combined	10%	75%

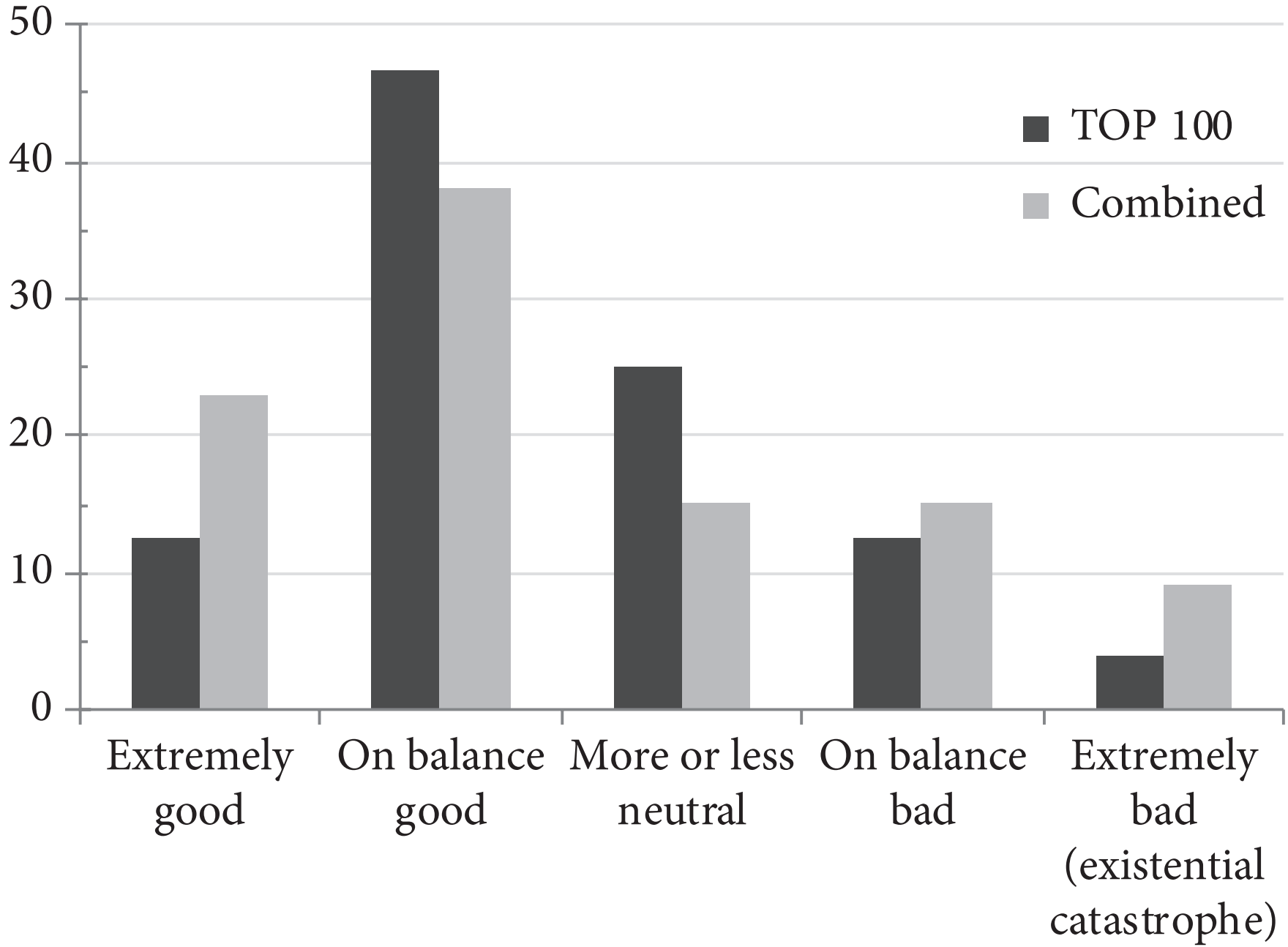

Figure 2 Overall long-term impact of HLMI.[83]

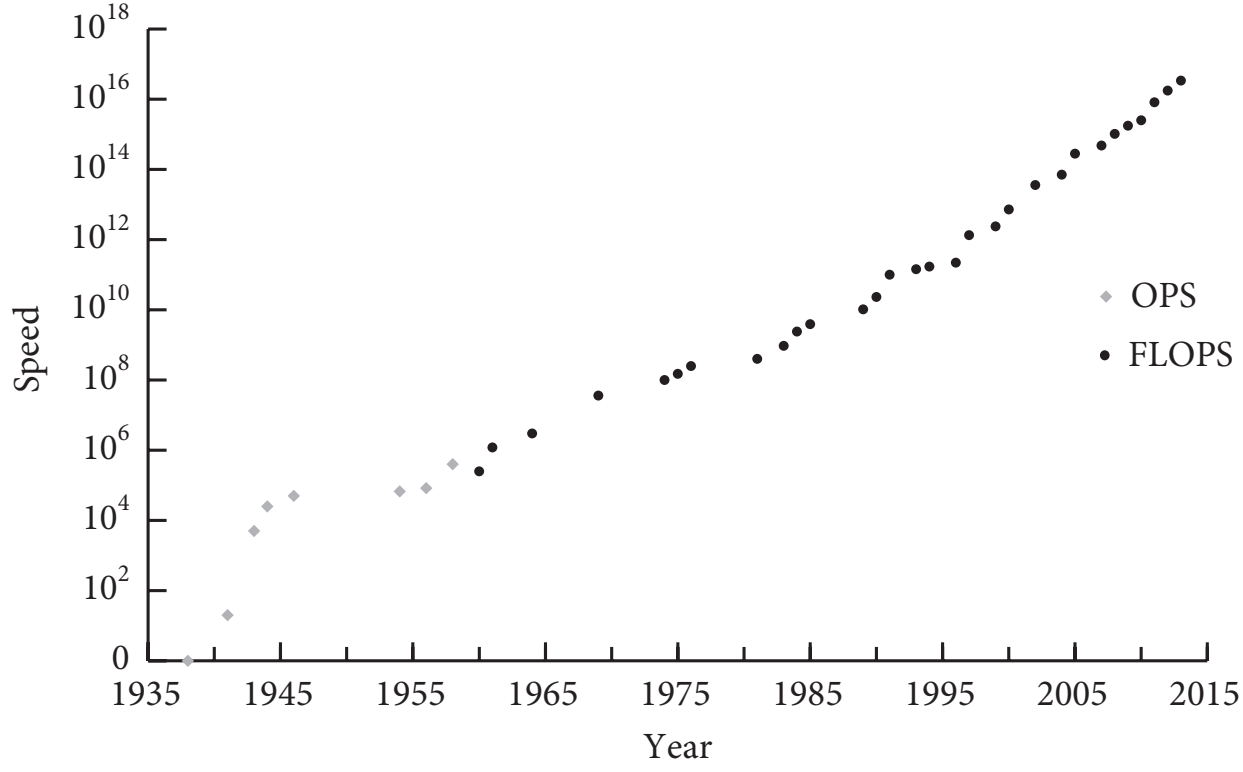

Figure 3 Supercomputer performance. In a narrow sense, "Moore's law" refers to the observation that the number of transistors on integrated circuits have for several decades doubled approximately every two years. However, the term is often used to refer to the more general observation that many performance metrics in computing technology have followed a similarly fast exponential trend. Here we plot peak speed of the world's fastest supercomputer as a function of time (on a logarithmic vertical scale). In recent years, growth in the serial speed of processors has stagnated, but increased use of parallelization has enabled the total number of computations performed to remain on the trend line.[16]

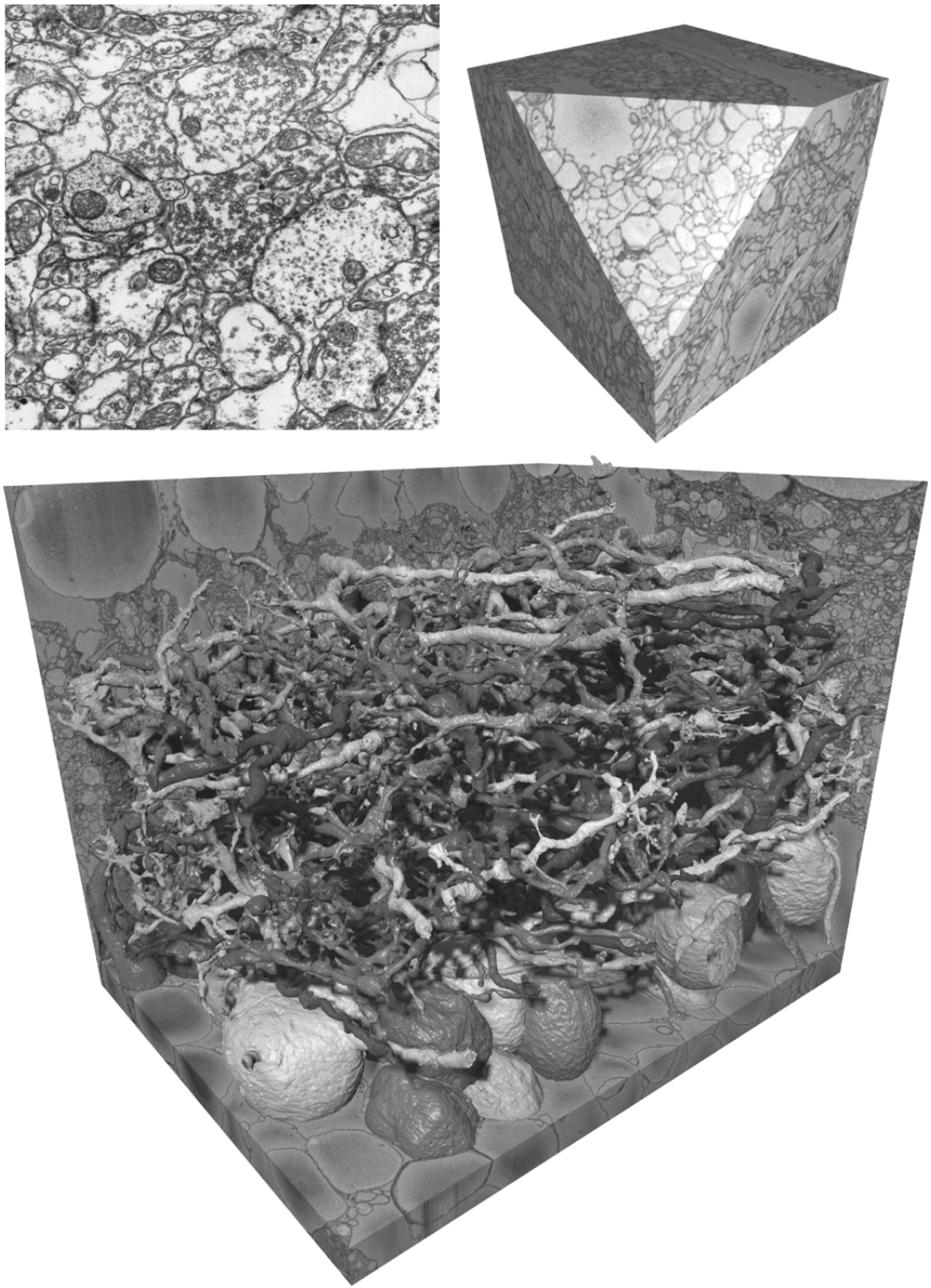

Figure 4 Reconstructing 3D neuroanatomy from electron microscope images. *Upper left*: A typical electron micrograph showing cross-sections of neuronal matter—dendrites and axons. *Upper right*: Volume image of rabbit retinal neural tissue acquired by serial block-face scanning electron microscopy.[21] Individual 2D images have been stacked into a cube (with a side of approximately 11 μm). *Bottom*: Reconstruction of a subset of the neuronal projections filling a volume of neuropil, generated by an automated segmentation algorithm.[22]

Table 4 ***Capabilities needed for whole brain emulation***

Scanning	Pre-processing/fixation		Preparing brains appropriately, retaining relevant microstructure and state
	Physical handling		Methods of manipulating fixed brains and tissue pieces before, during, and after scanning
	Imaging	Volume	Capability to scan entire brain volumes in reasonable time and expense
		Resolution	Scanning at sufficient resolution to enable reconstruction
		Functional information	Ability for scanning to detect the functionally relevant properties of tissue
Translation	Image processing	Geometric adjustment	Handling distortions due to scanning imperfections
		Data interpolation	Handling missing data
		Noise removal	Improving scan quality
		Tracing	Detecting structure and processing it into a consistent 3D model of the tissue
	Scan interpretation	Cell type identification	Identifying cell types
		Synapse identification	Identifying synapses and their connectivity
		Parameter estimation	Estimating functionally relevant parameters of cells, synapses, and other entities
		Databasing	Storing the resulting inventory in an efficient way
	Software model of neural system	Mathematical model	Model of entities and their behavior
		Efficient implementation	Implementation of model
Simulation	Storage		Storage of original model and current state
	Bandwidth		Efficient interprocessor communication
	CPU		Processor power to run simulation
	Body simulation		Simulation of body enabling interaction with virtual environment or actual environment via robot
	Environment simulation		Virtual environment for virtual body

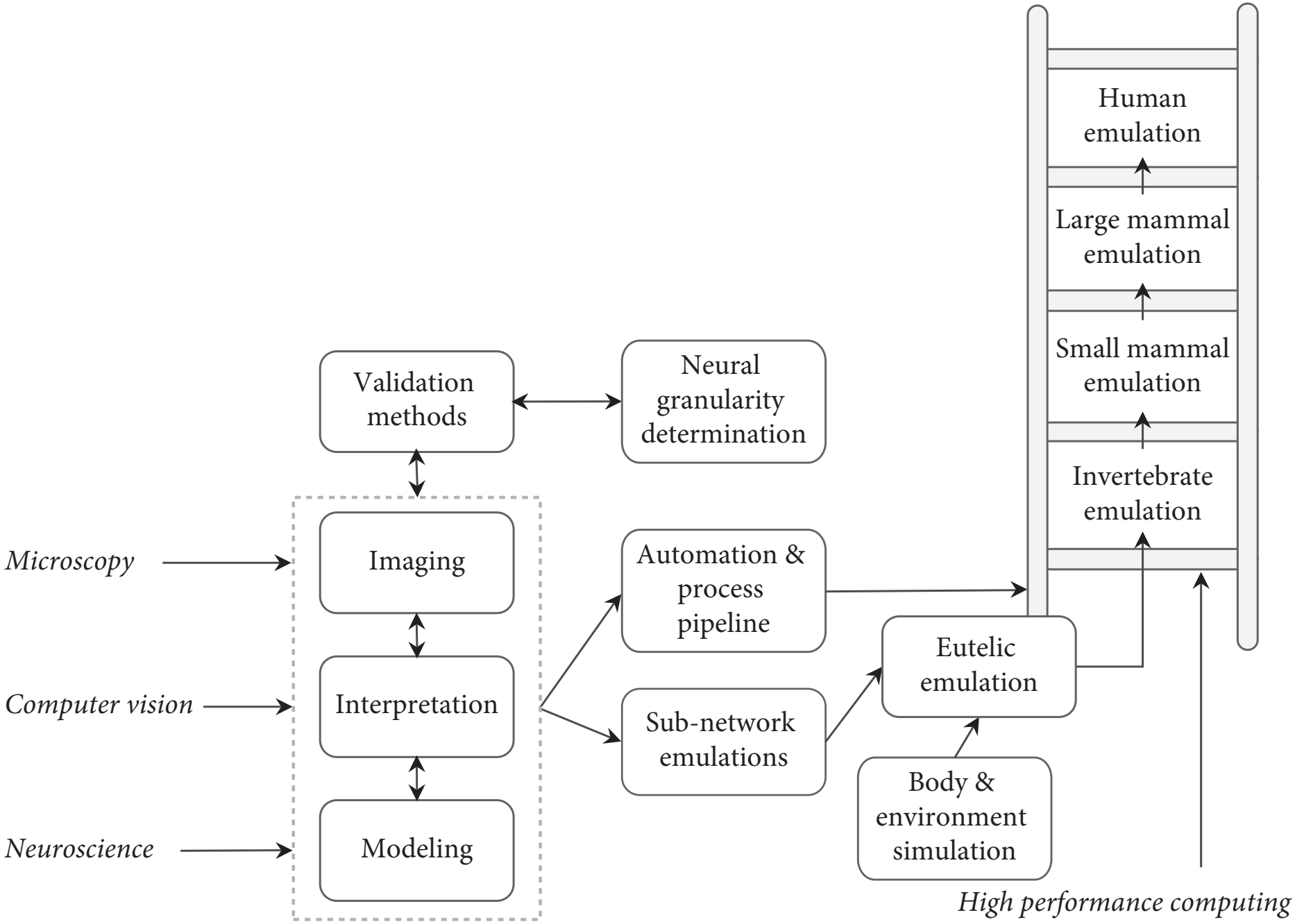

Figure 5 Whole brain emulation roadmap. Schematic of inputs, activities, and milestones.[28]

Table 5 ***Maximum IQ gains from selecting among a set of embryos***[43]

Selection	IQ points gained
1 in 2	4.2
1 in 10	11.5
1 in 100	18.8
1 in 1000	24.3
5 generations of 1 in 10	< 65 (b/c diminishing returns)
10 generations of 1 in 10	< 130 (b/c diminishing returns)
Cumulative limits (additive variants optimized for cognition)	100 + (< 300 (b/c diminishing returns))

Table 6 ***Possible impacts from genetic selection in different scenarios***[52]

Adoption / technology	"IVF+" Selection of 1 of 2 embryos [4 points]	"Aggressive IVF" Selection of 1 of 10 embryos [12 points]	"*In vitro* egg" Selection of 1 of 100 embryos [19 points]	"Iterated embryo selection" [100+ points]
"Marginal fertility practice" ~ 0.25% adoption	Socially negligible over one generation. Effects of social controversy more important than direct impacts.	Socially negligible over one generation. Effects of social controversy more important than direct impacts.	Enhanced contingent form noticeable minority in highly cognitively selective positions.	Selected dominate ranks of elite scientists, attorneys, physicians, engineers. Intellectual Renaissance?
"Elite advantage" 10% adoption	Slight cognitive impact in 1st generation, combines with selection for non-cognitive traits to perceptibly advantage a minority.	Large fraction of Harvard undergraduates enhanced. 2nd generation dominate cognitively demanding professions.	Selected dominate ranks of scientists, attorneys, physicians, engineers in 1st generation.	"Posthumanity"[53]
"New normal" > 90% adoption	Learning disability much less frequent among children. In 2nd generation, population above high IQ thresholds more than doubled.	Substantial growth in educational attainment, income. 2nd generation manyfold increase at right tail.	Raw IQs typical for eminent scientists 10+ times as common in 1st generation. Thousands of times in 2nd generation.	"Posthumanity"

Figure 6 Composite faces as a metaphor for spell-checked genomes. Each of the central pictures was produced by superimposing photographs of sixteen different individuals (residents of Tel Aviv). Composite faces are often judged to be more beautiful than any of the individual faces of which they are composed, as idiosyncratic imperfections are averaged out. Analogously, by removing individual mutations, proofread genomes may produce people closer to "Platonic ideals." Such individuals would not all be genetically identical, because many genes come in multiple equally functional alleles. Proofreading would only eliminate variance arising from deleterious mutations.[59]

Table 7 ***Some strategically significant technology races***

	United States	Soviet Union	United Kingdom	France	China	India	Israel	Pakistan	North Korea	South Africa
Fission bomb	1945	1949	1952	1960	1964	1974	1979?	1998	2006	1979?
Fusion bomb	1952	1953[11]	1957	1968	1967	1998	?	—	—	—
Satellite launch capability	1958	1957	1971	1965	1970	1980	1988	—	1998?[12]	—[13]
Human launch capability	1961	1961	—	—	2003	—	—	—	—	—
ICBM[14]	1959	1960	1968[15]	1985	1971	2012	2008	—[16]	2006	—[17]
MIRV[18]	1970	1975	1979	1985	2007	2014[19]	2008?			

Table 8 ***Superpowers: some strategically relevant tasks and corresponding skill sets***

Task	Skill set	Strategic relevance
Intelligence amplification	AI programming, cognitive enhancement research, social epistemology development, etc.	• System can bootstrap its intelligence
Strategizing	Strategic planning, forecasting, prioritizing, and analysis for optimizing chances of achieving distant goal	• Achieve distant goals • Overcome intelligent opposition
Social manipulation	Social and psychological modeling, manipulation, rhetoric persuasion	• Leverage external resources by recruiting human support • Enable a "boxed" AI to persuade its gatekeepers to let it out • Persuade states and organizations to adopt some course of action
Hacking	Finding and exploiting security flaws in computer systems	• AI can expropriate computational resources over the Internet • A boxed AI may exploit security holes to escape cybernetic confinement • Steal financial resources • Hijack infrastructure, military robots, etc.
Technology research	Design and modeling of advanced technologies (e.g. biotechnology, nanotechnology) and development paths	• Creation of powerful military force • Creation of surveillance system • Automated space colonization
Economic productivity	Various skills enabling economically productive intellectual work	• Generate wealth which can be used to buy influence, services, resources (including hardware), etc.

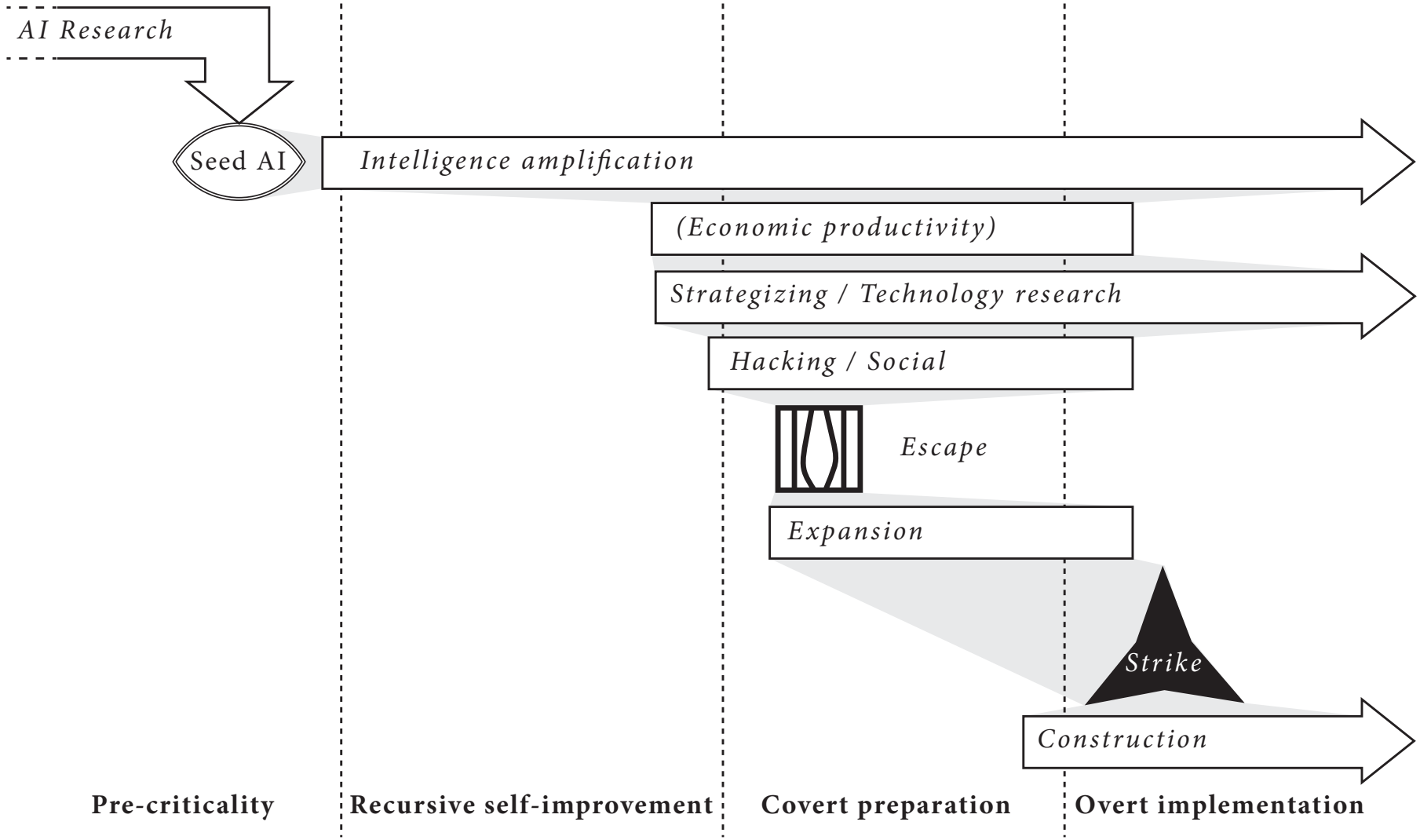

Figure 10 Phases in an AI takeover scenario.

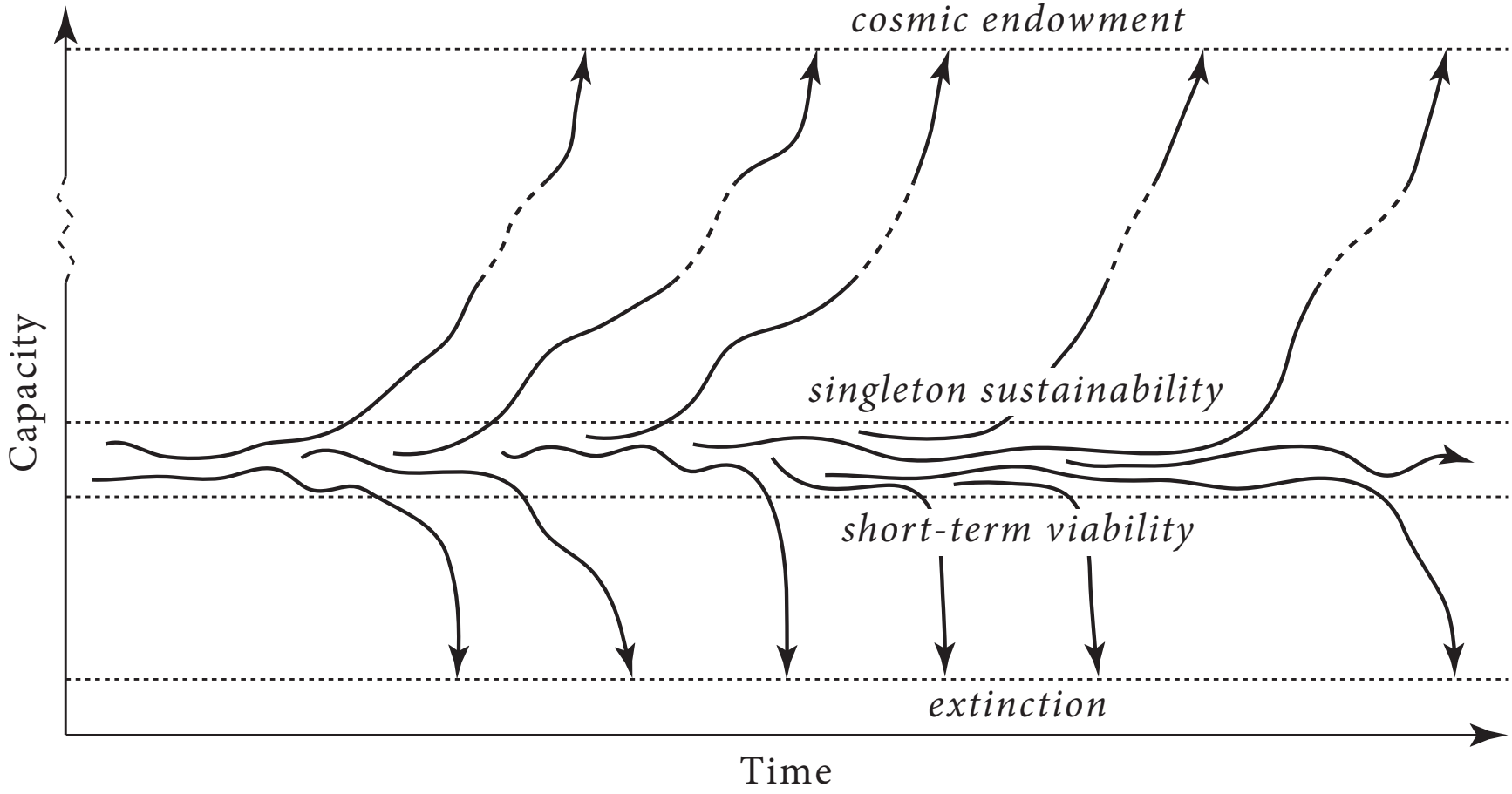

Figure 11 Schematic illustration of some possible trajectories for a hypothetical wise singleton. With a capability below the short-term viability threshold—for example, if population size is too small—a species tends to go extinct in short order (and remain extinct). At marginally higher levels of capability, various trajectories are possible: a singleton might be unlucky and go extinct or it might be lucky and attain a capability (e.g. population size, geographical dispersion, technological capacity) that crosses the wise-singleton sustainability threshold. Once above this threshold, a singleton will almost certainly continue to gain in capability until some extremely high capability level is attained. In this picture, there are two attractors: extinction and astronomical capability. Note that, for a wise singleton, the distance between the short-term viability threshold and the sustainability threshold may be rather small.[15]

Figure 12 Results of anthropomorphizing alien motivation. Least likely hypothesis: space aliens prefer blondes. More likely hypothesis: the illustrators succumbed to the "mind projection fallacy." Most likely hypothesis: the publisher wanted a cover that would entice the target demographic.

Table 9 ***Different kinds of tripwires***

Type of monitoring	Description and examples
Behavior	Detectors could be placed around a boxed AI to detect attempts to breach the containment. For example, detectors could intercept attempts at radio communication or at accessing internal computational resources intended to be off limits. An "Ethernet port of Eden" could be installed: an apparent connection to the internet that leads to a shutdown switch.
Ability	Automated capability testing could be performed at frequent intervals to determine the AI's skill in various domains. If either the rate of improvement is unexpectedly high or the AI attains a level of competence that brings it close to the potential danger zone, the system could be slowed down or paused, in order to allow the programmers to monitor further developments more closely.
Content	The AI's internal processes could be monitored for suspicious activity. For example, any change to the AI's representation of its final values might trigger an automatic shutdown and review. The AI's resource use and algorithmic methods may also be also be monitored for unexpected patterns. More ambitiously, the AI's beliefs, plans, subgoals, and justifications may be continuously scrutinized, to ensure they conform to the programmers' expectations. Such monitoring of the AI's intentional states would enable a particularly powerful form of content monitoring: the scanning of the AI's cognitive processes for a *conception of deception*—that vulnerable moment when an unfriendly AI first forms the intention to conceal its true intentions.[21] Content monitoring that requires that the AI's intentional states be transparent to the programmers or to an automatic monitoring mechanism may not be feasible for all kinds of AI architectures. (Some neural networks, for instance, are opaque, as they represent information holistically and in ways that do not necessarily match up with human concepts.) This may be a reason to avoid using such architectures.

Table 10 ***Control methods***

Capability control	
Boxing methods	The system is confined in such a way that it can affect the external world only through some restricted, pre-approved channel. Encompasses physical and informational containment methods.
Incentive methods	The system is placed within an environment that provides appropriate incentives. This could involve social integration into a world of similarly powerful entities. Another variation is the use of (cryptographic) reward tokens. "Anthropic capture" is also a very important possibility but one that involves esoteric considerations.
Stunting	Constraints are imposed on the cognitive capabilities of the system or its ability to affect key internal processes.
Tripwires	Diagnostic tests are performed on the system (possibly without its knowledge) and a mechanism shuts down the system if dangerous activity is detected.
Motivation selection	
Direct specification	The system is endowed with some directly specified motivation system, which might be consequentialist or involve following a set of rules.
Domesticity	A motivation system is designed to severely limit the scope of the agent's ambitions and activities.
Indirect normativity	Indirect normativity could involve rule-based or consequentialist principles, but is distinguished by its reliance on an indirect approach to specifying the rules that are to be followed or the values that are to be pursued.
Augmentation	One starts with a system that already has substantially human or benevolent motivations, and enhances its cognitive capacities to make it superintelligent.

Table 11 ***Features of different system castes***

Oracle	A question-answering system *Variations*: Domain-limited oracles (e.g. mathematics); output-restricted oracles (e.g. only yes/no/undecided answers, or probabilities); oracles that refuse to answer questions if they predict the consequences of answering would meet pre-specified "disaster criteria"; multiple oracles for peer review	• Boxing methods fully applicable • Domesticity fully applicable • Reduced need for AI to understand human intentions and interests (compared to genies and sovereigns) • Use of yes/no questions can obviate need for a metric of the "usefulness" or "informativeness" of answers • Source of great power (might give operator a decisive strategic advantage) • Limited protection against foolish use by operator • Untrustworthy oracles could be used to provide answers that are hard to find but easy to verify • Weak verification of answers may be possible through the use of multiple oracles
Genie	A command-executing system *Variations*: Genies using different "extrapolation distances" or degrees of following the spirit rather than letter of the command; domain-limited genies; genies-with-preview; genies that refuse to obey commands if they predict the consequences of obeying would meet pre-specified "disaster criteria"	• Boxing methods partially applicable (for spatially limited genies) • Domesticity partially applicable • Genie could offer a preview of salient aspects of expected outcomes • Genie could implement change in stages, with opportunity for review at each stage • Source of great power (might give operator a decisive strategic advantage) • Limited protection against foolish use by operator • Greater need for AI to understand human interests and intentions (compared to oracles)

Table 11 *Continued*

Sovereign	**A system designed for open-ended autonomous operation** *Variations*: Many possible motivation systems; possibility of using preview and "sponsor ratification" (to be discussed in Chapter 13)	• Boxing methods inapplicable • Most other capability control methods also inapplicable (except, possibly, social integration or anthropic capture) • Domesticity mostly inapplicable • Great need for AI to understand true human interests and intentions • Necessity of getting it right on the first try (though, to a possibly lesser extent, this is true for all castes) • Potentially a source of great power for sponsor, including decisive strategic advantage • Once activated, not vulnerable to hijacking by operator, and might be designed with some protection against foolish use • Can be used to implement "veil of ignorance" outcomes (cf. Chapter 13)
Tool	**A system not designed to exhibit goal-directed behavior**	• Boxing methods may be applicable, depending on the implementation • Powerful search processes would likely be involved in the development and operation of a machine superintelligence • Powerful search to find a solution meeting some formal criterion can produce solutions that meet the criterion in an unintended and dangerous way • Powerful search might involve secondary, internal search and planning processes that might find dangerous ways of executing the primary search process

Box 10 Formalizing value learning

Introducing some formal notation can help us see some things more clearly. However, readers who dislike formalism can skip this part.

Consider a simplified framework in which an agent interacts with its environment in a finite number of discrete cycles.[13] In cycle k, the agent performs action y_k, and then receives the percept x_k. The interaction history of an agent with lifespan m is a string $y_1x_1y_2x_2 \ldots y_mx_m$ (which we can abbreviate as $yx_{1:m}$ or $yx_{\leq m}$). In each cycle, the agent selects an action based on the percept sequence it has received to date.

Consider first a reinforcement learner. An optimal reinforcement learner (AI-RL) is one that maximizes expected future rewards. It obeys the equation[14]

$$y_k = \arg\max_{y_k} \sum_{x_k y x_{k+1:m}} (r_k + \ldots + r_m) P(yx_{\leq m} \mid yx_{<k} y_k).$$

The reward sequence $r_k, \ldots, r_m$ is implied by the percept sequence $x_{k:m}$, since the reward that the agent receives in a given cycle is part of the percept that the agent receives in that cycle.

As argued earlier, this kind of reinforcement learning is unsuitable in the present context because a sufficiently intelligent agent will realize that it could secure maximum reward if it were able to directly manipulate its reward signal (wireheading). For weak agents, this need not be a problem, since we can physically prevent them from tampering with their own reward channel. We can also control their environment so that they receive rewards only when they act in ways that are agreeable to us. But a reinforcement learner has a strong incentive to eliminate this artificial dependence of its rewards on our whims and wishes. Our relationship with a reinforcement learner is therefore fundamentally antagonistic. If the agent is strong, this spells danger.

Variations of the wireheading syndrome can also affect systems that do not seek an external sensory reward signal but whose goals are defined as the attainment of some internal state. For example, in so-called "actor–critic" systems, there is an actor module that selects actions in order to minimize the disapproval of a separate critic module that computes how far the agent's behavior falls short of a given performance measure. The problem with this setup is that the actor module may realize that it can minimize disapproval by modifying the critic or eliminating it altogether—much like a dictator who dissolves the parliament and nationalizes the press. For limited systems, the problem can be avoided simply by not giving the actor module any means of modifying the critic module. A sufficiently intelligent and resourceful actor module, however, could always gain access to the critic module (which, after all, is merely a physical process in some computer).[15]

Before we get to the value learner, let us consider as an intermediary step what has been called an observation-utility maximizer (AI-OUM). It is obtained

Box 10 *Continued*

by replacing the reward series ($r_k + \ldots + r_m$) in the AI-RL with a utility function that is allowed to depend on the entire future interaction history of the AI:

$$y_k = \arg\max_{y_k} \sum_{x_k y x_{k+1:m}} U(yx_{\leq m}) P(yx_{\leq m} \,|\, yx_{<k} y_k).$$

This formulation provides a way around the wireheading problem because a utility function defined over an entire interaction history could be designed to penalize interaction histories that show signs of self-deception (or of a failure on the part of the agent to invest sufficiently in obtaining an accurate view of reality).

The AI-OUM thus makes it possible *in principle* to circumvent the wireheading problem. Availing ourselves of this possibility, however, would require that we specify a suitable utility function over the class of possible interaction histories—a task that looks forbiddingly difficult.

It may be more natural to specify utility functions directly in terms of possible worlds (or properties of possible worlds, or theories about the world) rather than in terms of an agent's own interaction histories. If we use this approach, we could reformulate and simplify the AI-OUM optimality notion:

$$y = \arg\max_y \sum_w U(w) P(w \,|\, Ey).$$

Here, E is the total evidence available to the agent (at the time when it is making its decision), and U is a utility function that assigns utility to some class of possible worlds. The optimal agent chooses the act that maximizes expected utility.

An outstanding problem with these formulations is the difficulty of defining the utility function U. This, finally, returns us to the value-loading problem. To enable the utility function to be learned, we must expand our formalism to allow for uncertainty over utility functions. This can be done as follows (AI-VL):[16]

$$y = \arg\max_{y \in \mathbb{Y}} \sum_{w \in \mathbb{W}} P(w \,|\, Ey) \sum_{u \in \mathbb{U}} U(w) P(\mathcal{V}(U) \,|\, w).$$

Here, $\mathcal{V}(.)$ is a function from utility functions to propositions about utility functions. $\mathcal{V}(U)$ is the proposition that the utility function U satisfies the *value criterion* expressed by $\mathcal{V}$.[17]

To decide which action to perform, one could hence proceed as follows: First, compute the conditional probability of each possible world w (given available evidence and on the supposition that action y is to be performed). Second, for each possible utility function U, compute the conditional probability that U satisfies the value criterion $\mathcal{V}$ (conditional on w being the actual world). Third, for each possible utility function U, compute the utility of possible world w. Fourth, combine these quantities to compute the expected utility of action y. Fifth, repeat this procedure for each possible action, and perform the action found to have the highest

continued

Box 10 *Continued*

expected utility (using some arbitrary method to break ties). As described, this procedure—which involves giving explicit and separate consideration to each possible world—is, of course, wildly computationally intractable. The AI would have to use computational shortcuts that approximate this optimality notion.

The question, then, is how to define this value criterion $\mathcal{V}$.[18] Once the AI has an adequate representation of the value criterion, it could in principle use its general intelligence to gather information about which possible worlds are most likely to be the actual one. It could then apply the criterion, for each such plausible possible world *w*, to find out which utility function satisfies the criterion $\mathcal{V}$ in *w*. One can thus regard the AI-VL formula as a way of identifying and separating out this key challenge in the value learning approach—the challenge of how to represent $\mathcal{V}$. The formalism also brings to light a number of other issues (such as how to define $\mathbb{Y}$, $\mathbb{W}$, and $\mathbb{U}$) which would need to be resolved before the approach could be made to work.[19]

Table 12 ***Summary of value-loading techniques***

Explicit representation	May hold promise as a way of loading domesticity values. Does not seem promising as a way of loading more complex values.
Evolutionary selection	Less promising. Powerful search may find a design that satisfies the formal search criteria but not our intentions. Furthermore, if designs are evaluated by running them—including designs that do not even meet the formal criteria—a potentially grave additional danger is created. Evolution also makes it difficult to avoid massive mind crime, especially if one is aiming to fashion human-like minds.
Reinforcement learning	A range of different methods can be used to solve "reinforcement-learning problems," but they typically involve creating a system that seeks to maximize a reward signal. This has an inherent tendency to produce the wireheading failure mode when the system becomes more intelligent. Reinforcement learning therefore looks unpromising.
Value accretion	We humans acquire much of our specific goal content from our reactions to experience. While value accretion could in principle be used to create an agent with human motivations, the human value-accretion dispositions might be complex and difficult to replicate in a seed AI. A bad approximation may yield an AI that generalizes differently than humans do and therefore acquires unintended final goals. More research is needed to determine how difficult it would be to make value accretion work with sufficient precision.
Motivational scaffolding	It is too early to tell how difficult it would be to encourage a system to develop internal high-level representations that are transparent to humans (while keeping the system's capabilities below the dangerous level) and then to use those representations to design a new goal system. The approach might hold considerable promise. (However, as with any untested approach that would postpone much of the hard work on safety engineering until the development of human-level AI, one should be careful not to allow it to become an excuse for a lackadaisical attitude to the control problem in the interim.)

Table 12 *Continued*

Value learning	A potentially promising approach, but more research is needed to determine how difficult it would be to formally specify a reference that successfully points to the relevant external information about human value (and how difficult it would be to specify a correctness criterion for a utility function in terms of such a reference). Also worth exploring within the value learning category are proposals of the Hail Mary type or along the lines of Paul Christiano's construction (or other such shortcuts).
Emulation modulation	If machine intelligence is achieved via the emulation pathway, it would likely be possible to tweak motivations through the digital equivalent of drugs or by other means. Whether this would enable values to be loaded with sufficient precision to ensure safety even as the emulation is boosted to superintelligence is an open question. (Ethical constraints might also complicate developments in this direction.)
Institution design	Various strong methods of social control could be applied in an institution composed of emulations. In principle, social control methods could also be applied in an institution composed of artificial intelligences. Emulations have some properties that would make them easier to control via such methods, but also some properties that might make them harder to control than AIs. Institution design seems worthy of further exploration as a potential value-loading technique.

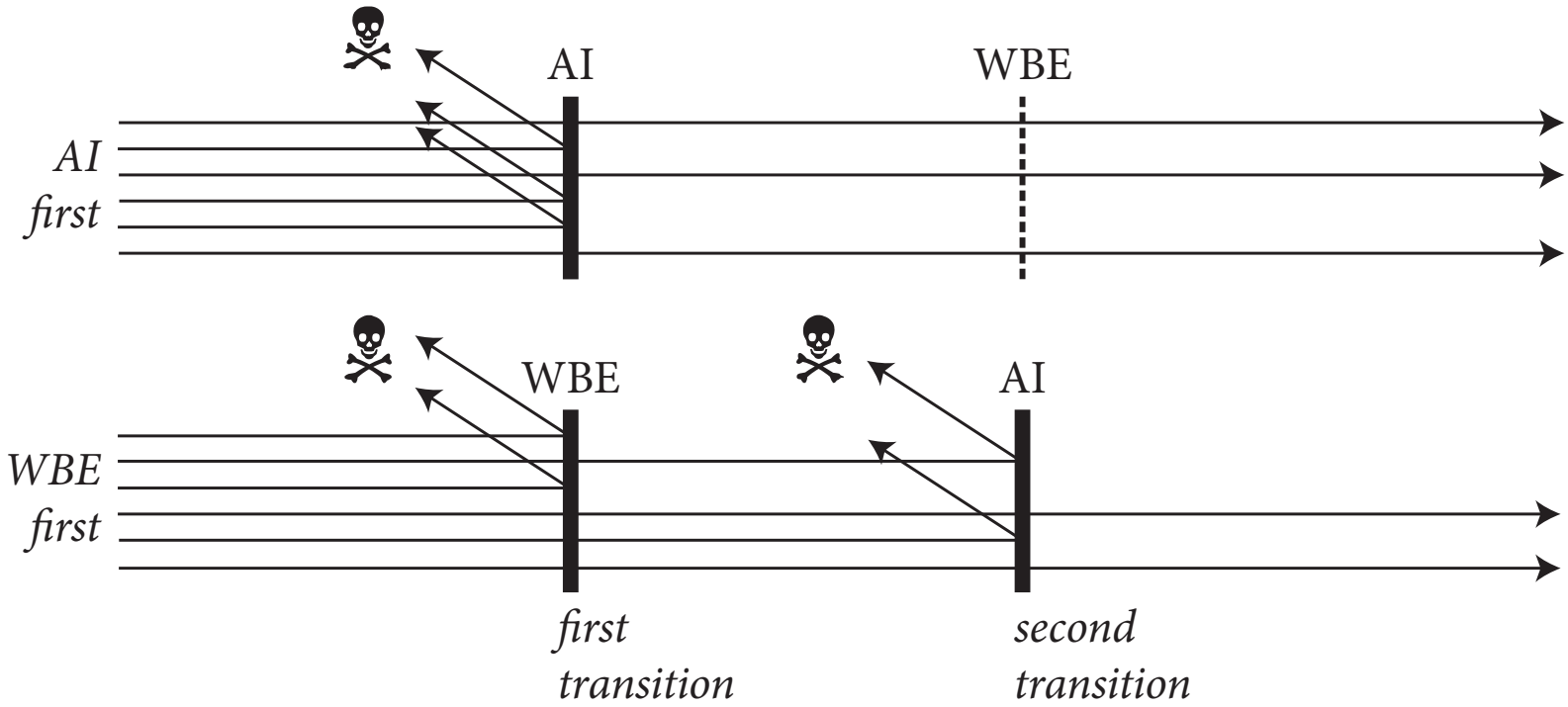

Figure 13 Artificial intelligence or whole brain emulation first? In an AI-first scenario, there is one transition that creates an existential risk. In a WBE-first scenario, there are two risky transitions, first the development of WBE and then the development of AI. The total existential risk along the WBE-first scenario is the sum of these. However, the risk of an AI transition might be lower if it occurs in a world where WBE has already been successfully introduced.

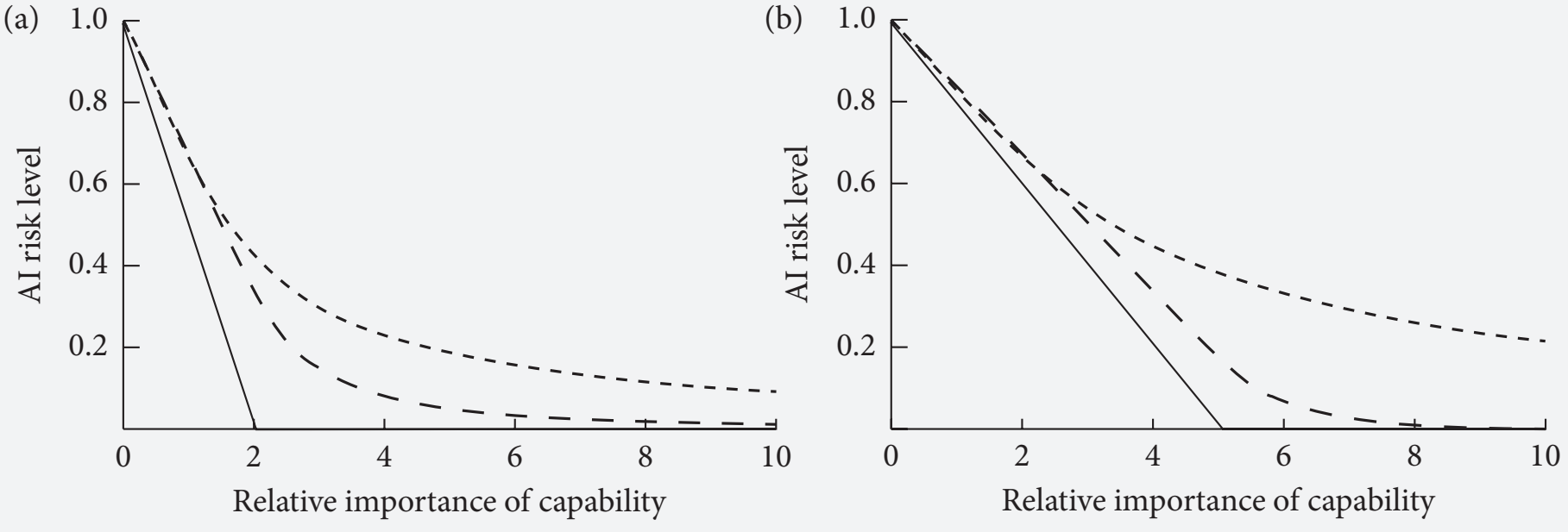

Figure 14 Risk levels in AI technology races. Levels of risk of dangerous AI in a simple model of a technology race involving either (a) two teams or (b) five teams, plotted against the relative importance of capability (as opposed to investment in safety) in determining which project wins the race. The graphs show three information-level scenarios: no capability information (straight), private capability information (dashed), and full capability information (dotted).

continued