
Human Amplification, Intelligent Agents, and the Aims of AI Research

Tushita Jha*
Mimir Center for Long Term Futures Research

Tom Everitt*
Google DeepMind

Alex Grzankowski
King's College London and Institute of Philosophy, School of Advanced Study, University of London

Abstract

As AI capabilities grow, the field faces a choice in how to define success. We argue that the primary objective of AI research should be Human Amplification – aiming for systems that augment human abilities and preserve human agency. While the current focus on intelligent agents drives significant progress, treating replication of human agency as the ultimate metric of success emphasizes delegation over augmentation. We discuss how this presents distinct challenges regarding control and power dynamics. By contrast, anchoring AI development in the service of human agency opens up a much broader design space that blurs the distinction between non-agentic tools and sophisticated delegates. This shift directs research toward cognitive extensions that ideally spark a virtuous cycle of recursive human amplification, where advances in AI further empower us to build safer and more beneficial systems.

Contents

1	Introduction	2
2	What is the Current Aim of AI Research?	4
3	Two Approaches to AI	5
4	Risks of the Iso-Synthetic Path	6
5	The Value of Human Agency	7
6	Recursive Human Amplification	8
7	Human Amplification in Practice	9
8	How Can We Steer AI Towards Human Amplification?	10
9	Conclusions	12
A	Frequently Asked Questions	20

*Joint first author

1 Introduction

A vision of artificial intelligence that centers on replicating human agency – building highly autonomous and generally intelligent systems that synthetically reconstruct key features of human capability – is both commonplace and alluring. But what if this very paradigm is leading us astray? Rather than empowering us, it risks trapping us in cycles of dependence and capability atrophy.

An implicit aspiration seems to be delegation: the creation of increasingly capable and autonomous artificial agents to take on tasks, solve problems, and perhaps even govern aspects of our world. Yet if we truly value agency (and we do!), shouldn't we prioritize increasing human agency rather than delegating it? In particular, shouldn't we aim to enhance our ability to understand the world, influence the world according to our values, and, when our behavior proves to be inadequate, enhance our ability to reflect upon it and modify it according to our reasoning?

A shift in aim away from replication towards amplification of human agency is not just a design preference – it's a rational response to the challenges we face presently and key uncertainties of the future. Amplified human agency is valuable for its own sake, but, importantly, it positions us to adapt, flourish, and thrive in ways that the focus on delegation and substitution is unlikely to achieve.

In this paper, we propose a radical reframing of AI's purpose: from replicating human agency to amplifying and extending it. This isn't merely a critique of current aspirations. It's a call to rethink the foundations of the field. By reorganizing our efforts around enhancing human capacities, we can open new paths towards mitigating risks of takeover, power concentration, and obsolescence that the current approach inherently invites. Our proposal is that rather than shaping agents into alignment, we should create human amplifiers from the very start.

We reflect upon how the idea of agency has taken a central role in the imagination and aspirations of the discipline of artificial intelligence. We propose a recentering of human agency as a normative standard for our shared endeavor:

The top-level aim of AI research should be to responsibly amplify human agency.

In particular, we distinguish between *iso-synthetic approaches* – which aim to build systems that replicate and surpass key features of human agency, from *co-synthetic approaches* – which aim to build systems that support and extend human agency, and argue for the importance of directing AI research toward co-synthetic approaches rather than iso-synthetic ones.

Note in particular that this change in aims does *not* imply that systems with autonomous agency must never be built. Our ask is much more modest than that. We only ask that the methodological focus on replicating human agency cease to be the top-level goal of AI research. When such systems are the best way to amplify human agency, we have no objection to their development. But given their risks and downsides, and the much broader space of possibilities, co-agentic alternatives can often be better options. In these cases, it would be a big mistake to dismiss these alternatives just because they don't formally conform to the agent paradigm.

Importantly, increases in human agency often enable further increases, as each increase enhances our ability to steer technological development towards further amplification. Thus, if we manage to build technology that leads to even modest increases in human agency, this increase might be enough to set off a virtuous cycle of *recursive human amplification* (Figure 1).

Finally, the expanded design space provides crucial flexibility for carrying out human amplification *responsibly*, with increases in agency widely and fairly distributed across groups and individuals, and with advances in destructive capacity not outpacing protective capabilities. Even when no single technical invention can achieve all these aspects at once, we are confident that for each aspect there are technical inventions that will help.

Key contributions. The key contribution of this paper is the distillation of central arguments for the just-mentioned change in aim of AI research. The central arguments are:

- the risks associated with pursuit of autonomous agents (Section 4),
- the value of human agency: it's intrinsically valuable to human well-being (Section 5), and provides safety through control and recursive human amplification (Section 6),
- the broad range of co-agentic possibilities (Section 7).

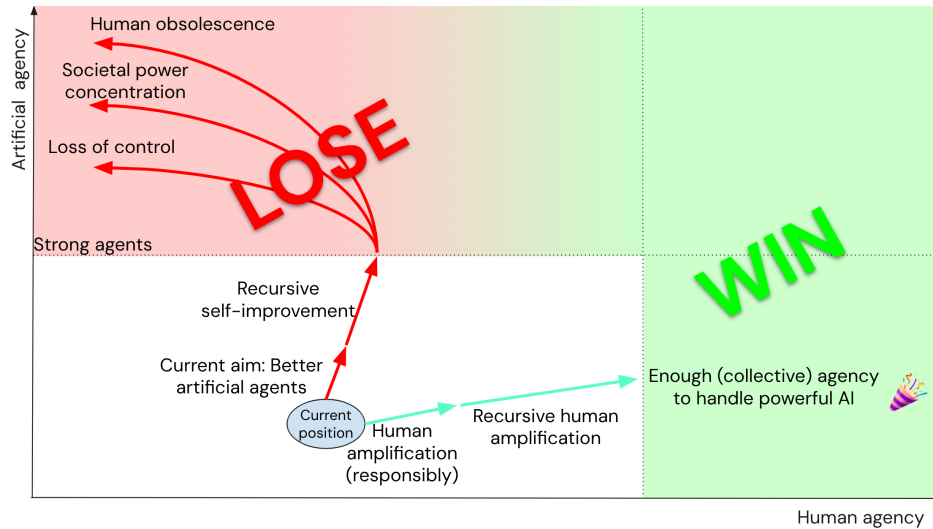


Figure 1: The strategic picture comparing an aim towards autonomous artificial agents with an aim towards amplification of human agency. Better artificial agents initially increase human agency. But strong versions also come with risks such as loss of control, societal power concentration and human obsolescence (Section 4). In the worst case, these can completely undermine human agency. In contrast, increases in human agency can lead to *recursive human amplification*, where increases in human agency improve our ability to develop technologies that further increase our agency.

To set the stage, we first review related work (just below), trace the current focus on artificial agents (Section 2), and outline two fundamentally different approaches to AI (Section 3). Finally, we suggest concrete research directions (Section 8) and offer some conclusions (Section 9). Frequently asked questions are answered in Appendix A.

Related work. Our paper builds on the tradition of intelligence augmentation, from cybernetics (Wiener, 1954; Ashby, 1956; Asaro, 2008; Lem, 1964) to human-computer interaction (Licklider, 1960; Engelbart, 1962; Rheingold, 2000; Maes, 2017; Nielsen and Matuschak, 2019). These ideas deserve renewed attention as AI begins to seriously reshape society.

A growing body of work critically examines the goals of AI research (Mitchell et al., 2025; Blili-Hamelin et al., 2024, 2025) and the risks from autonomous AI agents (Bengio et al., 2025; Aguirre, 2025; Chan et al., 2023; Cohen et al., 2022), including risks to human agency specifically (Kasirzadeh, 2025; Kulveit et al., 2025; Harari, 2024; Coeckelbergh, 2023, 2025). We share these concerns, and further elaborate an alternative research direction.

Several lines of work inform our positive proposal. Brynjolfsson (2022) frames the choice between automation and augmentation from an economic perspective. Capabiltarian ethics has long emphasized the deep value of human agency (Sen, 1985, 1993; Nussbaum, 2011; Robeyns and Byskov, 2020), and recent works consider its value in the context of AI (Mitelut et al., 2023; London and Heidari, 2024; Koralus, 2025) and the possibility of preserving human agency (Jin et al., 2025; Noller, 2024; Buterin, 2023). Like recursive alignment approaches (Leike et al., 2018; Christiano et al., 2018; Irving et al., 2018), we advocate using AI to increase human abilities – but emphasize that these increases have much broader value than better aligning powerful systems. Vaintroub and Cotton-Barratt (2025) argue for accelerating AI tools for epistemic improvement and coordination, a vision closely aligned with ours.

Finally, we draw on the literature discussing AI as extended cognition (Clark, 2025; White et al., 2025; Figà-Talamanca, 2026; Tong, 2026) and as global cognitive infrastructure (Heylighen, 2011; Heylighen and Lenartowicz, 2017; Last, 2020; Yager, 2024; Weyl et al., 2024). Some recent works have made practical progress on human amplification (Du et al., 2020; Ellis et al., 2025). Our position is similar to d/acc (Buterin, 2023), which argues for defensive, distributed, and democratic acceleration of human abilities. One might also call our approach h/acc, for human accelerationism.

2 What is the Current Aim of AI Research?

Artificial Intelligence as a research discipline spans academia, industry, and civil society. While its aims resist precise demarcation—given the plurality of interpretations, paradigms, and motivations throughout its nearly seven-decade history—reflecting on these aims remains crucial for understanding the visions and grand projects to which individual research contributes.

In its early days, AI research was often understood as a scientific inquiry into understanding intelligence in computational terms. The Dartmouth Conjecture from the field’s founding 1956 workshop captured this spirit:

"Every aspect of learning and any other feature of intelligence can be so precisely understood that a machine can be made to replicate it." (McCarthy et al., 1955)

The aim of this understanding differed, ranging from reproducing the generators of human-like cognition² ³ to the materialization of idealized reasoning⁴⁵. Similarly varied were the practical purposes for the artifacts this understanding produced: automation of cognitive labour (Lighthill, 1973; Simon, 1983; Nilsson, 2005), universal organization of information and knowledge production (Licklider, 1965; Davis et al., 1993; Feigenbaum, 1992; Heylighen and Lenartowicz, 2017; Lenat, 1995), and efficient administration (Simon, 1965; Vinuesa et al., 2020; Cockshott and Cottrell, 1993; Bullock et al., 2025; Parson, 2020).

Central to these practical ends was the ambitious conception of general-purpose and complex problem solving competence⁶ (Solomonoff et al., 2006; Newell and Simon, 1976; McCarthy, 1973; Nilsson, 1971; Newell and Simon, 1972). However, by the 1990s this pursuit of general problem solving had evolved into a focus on building rational agents, in what came to be known as the Intelligent Agents paradigm. In their field-defining textbook, Russell and Norvig (1995) write:

"AI is the field dedicated to theory and design of agents."

Russell and Norvig (1995) use a rather modest definition of agent. However, for many researchers, the Intelligent Agent paradigm became alluring precisely because it suggested the possibility of building systems with *autonomous agency* able to independently achieve their goals, thus replicating key features of human agency (Kurzweil et al., 1990; Sloman, 1993; Wooldridge and Jennings, 1995; Rao and Georgeff, 1995; Parkes and Wellman, 2015). In this paper, we take *agency* to mean the range of desired, real-world outcomes a system can independently achieve.⁷

The appeal of the Intelligent Agents paradigm comes from mainly two sources: functional compactness and practical utility. Functional compactness was the main reason rational agency served as a crucial backdrop to the Artificial General Intelligence movement, which gained traction in the early 2000s. The functional character of rational agency allowed the articulation of general problem solving that could be perfected to surpass human irrationalities and limitations (Goertzel, 2001; Solomonoff et al., 2006). These functional motivations played a significant role in the emergence of formal idealizations (Hutter, 2005; Friston et al., 2010), prognostic concerns (Good, 1966; Vinge, 1993; Joy, 2000; Bostrom, 2003), diagnostic interventions (Anderson and Anderson, 2007; Allen and Varner, 2000; Floridi and Sanders, 2004; Soares and Fallenstein, 2014), and practical criticisms (Suchman,

²"[The field of AI] aims at understanding the complex information processes that underlie man’s ability to solve problems, learn, adapt, and create." (Newell, 1963)

³Simon (1983) uses the term *cognitive simulation* for this interpretation of the field of AI.

⁴"[The aim of AI is] studying the structure of information and the structure of problem solving processes . . . independently of its realization in animals or humans." (McCarthy, 1973)

⁵Russell and Norvig (2010, p. 17) suggest the name *computational rationality* for this interpretation.

⁶Solomonoff et al. (2006) articulates this as: "I am interested in creating a machine that can work very difficult problems much better and/or faster than humans can – and this machine should be embodied in a technology to which Moore’s Law applies. I would like it to give a better understanding of the relation of quantum mechanics to general relativity. I would like it to discover cures for cancer and AIDS. I would like it to find some very good high temperature superconductors. I would not be disappointed if it were unable to pass itself off as a rock star."

⁷The wider the range, the greater the agency. Google Maps has limited agency; LLM agents somewhat more; human collectives more yet (List and Pettit, 2011). See also (Kasirzadeh and Gabriel, 2025; Dung, 2025; Bent, 2025; Floridi and Sanders, 2004; Franklin and Graesser, 1996) for various characterizations.

1987; Maes, 1990; Stanley and Lehman, 2015), as well in the eventual rise of reinforcement learning as a popular paradigm for pursuing AGI (Sutton et al., 1999).

A somewhat independent motivation for intelligent agents is the practical utility for automating complex economic roles. As Herbert Simon said in 1983, “AI is directed toward getting computers to be smart and do smart things so that human beings don’t have to do them”. This motivation emphasizes autonomy and robust capabilities that enable us to delegate cognitive labour to AI systems. Today, AI agents of practical utility are still shaping imaginations and strategies around AI commercialization (Strange and da Costa, 2024; BCG, 2024; Sukharevsky et al., 2025; Holmes, 2025). In fact, several critics of the current paradigm have gone on to propose research agendas to overcome what they perceive as inadequate agentic properties of current generative AI systems (LeCun, 2022; Kudithipudi et al., 2022; McShane et al., 2024; Heins et al., 2025).

3 Two Approaches to AI

The trajectory traced above reflects one particular relationship between AI and human agency – but it is not the only one. Agency can serve either as a template to be replicated or as a capacity to be extended. This choice carries profound methodological implications by setting the evaluative standards by which technical progress is assessed.

The first option, which we term the **iso-synthetic approach**, proceeds by abstracting and idealizing properties from natural intelligence – particularly human agency – and searches for artificial systems that match or exceed it⁸. The Intelligent Agents paradigm discussed in Section 2 exemplifies the iso-synthetic orientation: here, agency serves as a template to be replicated, with the aspiration toward completely autonomous agents. Indeed, Legg and Hutter (2007) define intelligence – the very metric of success for AI – as:

“Intelligence measures an agent’s ability to achieve goals in a wide range of environments.”

An alternative, **co-synthetic approach** is revealed by dualizing the Legg-Hutter definition:

“(the measure of) an environment’s ability to support a wide range of agents in achieving their goals.”

While the Legg-Hutter definition asks “who can win no matter where you drop them,” this dualized definition asks “what ensures that whoever you drop there, they can win?”. In other words, rather than abstracting away from natural intelligence to build standalone replacements, the co-synthetic approach aims to integrate artificial systems with human capabilities to amplify what humans can understand, decide, and accomplish.⁹

Importantly, the co-synthetic approach is not committed to AI with high autonomous agency. The central concern is instead what we might call *co-agency*: when integrated with a human agent, to what extent does an AI system expand the human’s capacity to understand their situation, deliberate about their values, and influence the world according to their authentic purposes?

This opens a considerably wider design space. Systems with very little agency of their own – a calculator, search engine, or decision support tool – can significantly amplify human agency; so too can systems with greater autonomous capabilities, such as research assistants or planning aids, provided their design remains oriented toward supporting rather than supplanting human judgment (further discussed in Section 7).

⁸Parkes and Wellman (2015) articulate this trajectory explicitly: “AI strives to construct—out of silicon (or whatever) and information—a synthetic homo economicus, perhaps more accurately termed *machina economicus*.” The genealogy is an instructive illustration of what we call iso-synthetic approach: *homo economicus* emerged as an idealized reduction of human economic behavior; *machina economicus* represents a transformation of the idealization into an end of AI research.

⁹This distinction echoes a fundamental divergence in biological strategy – between evolving a new organism optimized for independent survival across environments, and constructing an enriched ecological niche that amplifies what existing organisms can do (Figà-Talamanca, 2026; White et al., 2025). Where iso-synthetic approaches asymptote toward perfected synthetic organisms – autonomous entities with coherent goals, self-preservation, and general capabilities – co-synthetic approaches lean toward mind-extending niche construction.

4 Risks of the Iso-Synthetic Path

Many of the concerns raised around AI development (Chan et al., 2023), such as power concentration, human displacement, and loss of control, are primarily associated with the iso-synthetic approach.

4.1 Technological Challenges

Misalignment risks. Any agential system capable of carrying out complex economic tasks must be able to reason instrumentally, and able to predict and plan. However, such capabilities can lead to influence-seeking behavior: the agential system might “realize” that with more resources or with humans more friendly towards it, it has a higher chance of achieving its goal. Such reasoning can lead it to independently develop strategies for acquiring resources or manipulating humans, even if its developers did not intend it to (Omohundro, 2018; Bostrom, 2014; Carlsmith, 2022; Hardt and Mendler-Dünnér, 2023). Such divergences between intended behavior and actual behavior are commonly referred to as *misalignment problems*.

Hardness of alignment. Appropriately directing agential cognition poses several challenges. First, direction toward a goal can easily be conflated with either a nearby goal (goal misgeneralization; Shah et al., 2022; Di Langosco et al., 2022) or proxy measures (e.g. reward hacking; Everitt et al., 2021). In particular, certain classes of goal representations might be favored in either design or training, due to quantifiability bias or simplicity bias. This can bottleneck the autonomous agent’s ability to internalize complex objectives. Second, without grasping the context-sensitive and thick dimensions of human values (Foster, 2023; Nelson, 2023; Zhi-Xuan et al., 2024), even well-intentioned AI systems will fail to benefit us in practice. Furthermore, noticing such misalignment before harm has occurred can be hard as intelligent agents might deceive us (Carranza et al., 2023; Sharkey, 2022). Trustworthy autonomous agents also require appropriate forms of robustness and coherence, otherwise they may fail to pursue the same goal in a sensible way as they evolve and encounter new situations (Freiesleben and Grote, 2023; Sohl-Dickstein, 2023; Macmillan-Scott and Musolesi, 2025).

Co-synthetic advantages. Rather than abdicating control to artificial beings, whose values we must take great care to get right, the co-synthetic approach instead primarily focuses on cognitive tools and infrastructure, i.e. technologies that are unable to function independently without humans in the loop. Alignment is still needed, in the sense that we need to make sure that tools and delegates do the job as intended. But the stakes will typically be much lower.

4.2 Sociotechnical Challenges

Power concentration. Drago and Laine (2025) express concern that autonomous AI could fundamentally alter traditional avenues of human influence, potentially undermining the leverage that people hold as workers and as citizens exercising collective action. While a risk to take seriously, AI can also distribute power – for instance, by helping individuals better represent their interests (Kapoor et al., 2025). Co-synthetic approaches may be better suited for such broad empowerment: as cognitive tools, they integrate more naturally into the workflows of a wider range of people; by aligning with users’ authentic purposes, they can amplify the voices of the less powerful while channeling the influence of the more powerful through deliberative processes; and as public infrastructure, they can strengthen rather than bypass the institutional checks that democratic societies depend on (Weyl et al., 2024; Tsai et al., 2024; Hsu et al., 2022; Trask et al., 2020).

Displacement. A related concern involves the iso-synthetic focus on delegation. While *delegation* implies deliberate, voluntary choices, the reality may be more constrained. Human competencies are formed within complex ecosystems; relying on automation for core cognitive tasks can lead to enfeeblement, prompting further reliance. For example, delegating analytical writing may enhance short-term productivity while risking long-term reductions in critical thinking skills (Mollick, 2024). If the adoption of these systems is driven primarily by systemic economic pressures rather than individual agency, the transition risks becoming a process of involuntary displacement rather than voluntary assistance (Kulveit et al., 2025; Kasirzadeh, 2025; Shao et al., 2025). Furthermore, over-reliance on AI-curated environments could reduce the deep reflection necessary to conceive meaningful pursuits, a vulnerability already observed in broader shifts in digital media consumption (Hou et al., 2019).

In contrast, co-synthetic approaches aim to enhance everyone's agency, hopefully enabling everyone to contribute for the foreseeable future, and enabling people to defend their interests if AI eventually becomes so competent that even amplified humans have little left to contribute.

4.3 Solved Worlds

What kind of world does the iso-synthetic approach lead to? Several AI thinkers argue that the creation of general, fully autonomous AI agents will lead to a much reduced role for humans. For example, Bostrom (2012), echoing Good (1966), claims that "Machine intelligence is the last invention mankind ever needs to make." Taken to its logical conclusion, the iso-synthetic path thus leads to a *solved world* (Bostrom, 2024), where humans never need to do anything for instrumental reasons. While solved worlds are welcomed by some (Amodei, 2024; Bostrom, 2024), worlds where AI are doing things for us seem far less attractive to us than the corresponding co-synthetic vision, where *humans are doing ambitious things with AI*. But why exactly are solved worlds unattractive? The answer lies in the essential role that agency plays in a good human life.

5 The Value of Human Agency

The problems outlined in the preceding section largely derive from prioritizing AI agency over human agency. In fact, human agency is not merely instrumentally useful – it is constitutive of a good human life. Here we outline three reasons why.

Well-being. A sense of agency is crucial to human well-being (Crisp, 2021). For example, self-determination theory has conducted a number of studies across different cultures, all pointing to connections between agency (autonomy, effectiveness) and well-being (Ryan and Deci, 2024). Relatedly, (Lambert, 2006) have found that doing tasks with our hands, such as crafts and cooking, where we exhibit agency in the perhaps most visceral sense, protects against depression.

Self-creation. Part of the reason that agency is important for well-being may be that the actions that we take – and the reasons that we give for them, to ourselves and to others – are important for creating our identity and our values (Korsgaard, 2009). What Chang (2017) calls *hard choices* may be particularly important: "When our given reasons are on a par, as they are in hard choices, we have the normative power to create will-based reasons for ourselves to choose one alternative over the other. It is not facts beyond our agency that determine whether we should lead this kind of life rather than that, but us. In this way, we make ourselves who we are through the exercise of our normative powers in hard choices." Related arguments have been made by Sartre (1948) ("existence precedes essence"), and are supported by narrative theories of identity (McAdams, 2001).

Axiological uncertainty. Another, and perhaps more obvious reason agency is valuable, is that agency gives us control, and lets us shape our environment according to our values. Maintaining, and ideally enhancing, this control is essential. We should not give it up to a superior being – be it a state or an AI – even if it claims to understand our needs and preferences. Why? Because it's hard to know what we will want tomorrow, and harder still when looking far beyond tomorrow. Indeed, philosophers and moral experts are yet to agree on even the shape of the answer to questions of what's a good action and what's a good life. To secure a good future, we need to ensure both that we retain the options to influence (Mill, 1859; MacAskill et al., 2020), as well as the moral growth and understanding enabling us to make good choices. Such growth likely requires us to regularly practice our agency (cf. self-creation just above).

Collective agency. These arguments show that a good future requires us to retain individual agency. What about collective agency, such as the agency of communities, societies, and humanity as a whole (List and Pettit, 2011)? Collective agency is valuable for roughly analogous reasons. Axiological uncertainty is as much a reason for groups as for individuals to retain their agency, as it's often even harder for groups than for individuals to articulate their preferences to the satisfaction of all their members. Further, it is only by taking responsibility together that we can turn a group into a collective agent, mirroring the self-creating aspect of agency (Mill, 1865; Pateman, 1975; Dewey, 1927). Finally, well-functioning agentic collectives are fundamentally important to individual well-being, though of course the agency of the group must not be so strong as to suppress individual forms of agency

(Arendt, 1958)¹⁰. For these reasons, it would likely be a bad idea to replace the human institutions currently governing us, with an artificial agent, regardless how rational and value-aligned (Bostrom, 2014). In Sections 7.2 and 8.1 we sketch some possibilities for how AI can instead help facilitate collective agency through shared representation and reasoning. And it is worth noting that individual agency correctly construed contributes to collective agency: cooperating with a group is often the best way to achieve one's aims.

6 Recursive Human Amplification

Having argued that human agency is necessary, we now argue that its amplification is also nearly sufficient for a desirable future. Through a mechanism we call *recursive human amplification*, even modest initial gains in human agency can compound: each increase enhances our ability to produce further increases. This process, if sustained, can address the concerns raised in Section 4.

As defined in Section 2, agency measures the ability to achieve a wide range of desired outcomes. We currently have at least some agency, both as individuals and as collectives, and we can use it to try to increase our agency further. If each round of improvement enhances our capacity for the next, the result is a compounding process of recursive human amplification.

Absent external perturbations and fundamental limits to our ability to amplify ourselves, recursive human amplification may lead to an exponential increase in human agency. While it's not hard to think of external perturbations that could set us back (war, pandemics, political instability, natural catastrophes, detrimental new technologies, ...), the more agency we have, the better positioned we'll be to deal with these challenges. For example, an earthquake causes less disruption to a human society amplified with technology to predict it, mitigate its effects, and to recover afterwards (e.g. through more stable houses and machines to clear rubble). We also see no obvious upper bound to human amplification: as long as humans stay in charge of hard choices and high-level decisions, there is no obvious upper limit to how much our agency can benefit from external decision support and execution.

More concretely, AI has the potential to amplify our agency in several ways, from helping us understand the world and act on that understanding, to enhancing collective deliberation and decision-making (see Section 7 for concrete examples). It can also help us understand our authentic¹¹ preferences, e.g. by encouraging reflection over mindless entertainment, or by creating an external representation of our accumulated thoughts and insights.

Gains along different dimensions can reinforce each other: better understanding of the world supports better deliberation, which in turn supports more effective action – each of which creates the means for further improvement. At the same time, responsible amplification requires attention to how these gains are distributed across dimensions, individuals and groups. An increase in influence without a commensurate increase in wisdom can be detrimental, as can processes that benefit the agency of some at the cost of others. But ultimately the tool we have at our disposal to address these challenges is our ability to understand and to influence, i.e. our agency. So these caveats are not arguments to hold back, but arguments to move forward with intention and care.

Given the mounting challenges, it's essential that we do not dither, and instead deliberately accelerate the process of recursive human amplification. If we succeed, then by the time the technology for fully autonomous AI agents arrives, we will have the agency to deploy them safely – or to forgo them entirely, having already secured the benefits they were meant to provide (Figure 1).

¹⁰This highlights the importance of institutional dimensions that are constitutively crucial for *responsible* human amplification – the recognition of collective agency must navigate between the Scylla of individuals failing to gain from coordination, and the Charybdis of individual agency being stymied by the group. As co-synthetic approaches expand the space of possible coordination mechanisms, they may protect and enhance "cooperation across difference" (Weyl et al., 2024).

¹¹Humans are somewhat manipulable, which means adversarial actors can sometimes override our more authentic desires. Such manipulation is undermining our agency, as it means that we no longer pursue (and therefore are unlikely to achieve) our truly desired outcomes. With *authentic* preferences, agency, etc. we emphasize that the desires must be the result of appropriate reflection in a supportive environment, rather than the results of whims or manipulation (Prunkl, 2022; Christman, 2009, p. 155).

7 Human Amplification in Practice

What does it mean to focus AI research on human amplification? Appropriate artificial agents can certainly serve this purpose – functioning as assistants (such as co-pilots) or bounded-scope delegates. However, the risks discussed in Section 4 suggest wisdom in exploring form factors beyond agency altogether.

The shift from iso-synthetic to co-synthetic approaches fundamentally expands the design space. When the yardstick is no longer “how well does this system replicate agency” but rather “how well does this system extend human agency,” a much broader range of possibilities becomes apparent – possibilities that might otherwise remain unexplored or dismissed simply because they don’t conform to the agent paradigm.

7.1 Beyond the Necessity of Iso-Synthetic Approaches

A closer examination of the Intelligent Agents paradigm reveals a powerful motivational assumption: the equating of intelligence with agency. As discussed in Section 3, this assumption is exemplified by the Legg-Hutter (2007) definition, which treats intelligence as fundamentally the property of agents pursuing goals across environments. Such a conception can often carry an assumption that intellectual faculties (such as planning, abstract reasoning, learning, etc.) only truly constitute intelligence when they belong to a coherent agent-like whole. When intelligence is defined this way, the field naturally gravitates toward building systems with greater agentic capabilities. This equation may inadvertently restrict our attention to a narrower design space than necessary.¹²

Large language models (LLMs) suggest there might be alternative ways of thinking about intelligence. LLMs learn richly processual approximations of human reasoning – producing deduction chains, analogical leaps, and style transfers. These reasoning mechanisms can be functionally redeployed across myriad interfaces: chat co-pilots, theorem provers, anomaly detectors, and code synthesizers. Although some LLMs are becoming increasingly agentic, with the ability to do independent web searches and tool calls, “vanilla” LLMs still suggest the distinct possibility of non-agentic intelligence – cognitive processes that need not belong to a coherent agent-like whole.

Rather than agents, AI may be better understood as *reasoning engines*, i.e. as versatile sources of cognitive power that can be harnessed in countless ways. Just as the steam engine generates force that can be applied across unimagined domains, AI produces “reasoning force” that can be applied to many problems¹³. The concept of an *Intelligent Agent* thus represents a contingent product choice – a particular implementation strategy – not an inevitable endpoint.

Importantly, this expanded design space does not exclude systems with significant autonomous capabilities. Even highly autonomous systems can amplify human agency, provided their directive is to enable humans to pursue their goals rather than to achieve goals on their behalf – an orientation aligned with the capabilities approach in welfare economics (Robeyns and Byskov, 2020; London and Heidari, 2024).

However, as Reinhardt (2025) writes about automated manufacturing: “[by the time we have good] humanoid robots, we will also [have] task-specific hardware . . . that can do [the same things] cheaper, faster, and better.” The same is likely true also for more general AI agents. By the time we have developed the technology for full AI agents (with self-awareness, legal personhood, and ownership of capital and other resources), we’ll also have the technology for cognitive infrastructure and advanced tools, that can help humans understand their world and their values, and to execute ambitious plans. While tool and agent are often positioned as distinct capability categories, increasing interactivity and

¹²A shift towards alternative characterizations of intelligence would also align with established perspectives in psychology, where concepts like ‘intellectual functions,’ ‘expertise,’ and ‘reasoning’ – the very features we aim to emulate in artificial systems – are defined by their cognitive mechanics rather than their agential purposes (Bermúdez, 2014).

¹³The reasoning force conception may prove more compatible with reason-responsive theories of action (Fischer and Ravizza, 1998; Scanlon, 1998; Dancy, 2000; Setiya, 2007; Korsgaard, 2009) than with the rational agent theories (Russell, 2016; Shah, 2018; Sparks and Wright, 2025) of action. The former treat action as emerging from ongoing responsiveness to normative considerations that may remain contestable and revisable (Raz, 1999), while the latter models action through the agent’s predetermined preference structures (Zhi-Xuan et al., 2024).

adaptivity of advanced tools may blur this boundary¹⁴. Just as Reinhardt (2025) observes, “there’s a reason we don’t pull cars with mechanical horses” – the best way to achieve human amplification need not rest upon the imitation of human agents.

7.2 Concrete Visions for Co-Synthetic Approaches

Consider a future in which general reasoning abilities are cheaply accessible across society, like a public utility. Just as electricity was shaped into a wide range of appliances – many of them previously unimaginable – these reasoning abilities have been turned into well-defined cognitive appliances serving countless purposes. Institutions and norms have evolved around them, minimizing negative externalities and dangerous uses. Such cultural evolution was itself facilitated by the application of reasoning abilities to democratic deliberation, coordination, and wisdom cultivation. Researchers – some of them automated – continue to improve the quality of the utility. Crucially, its public generators lack any coherent goals or autonomous volition. Over time, this progress enables humans to become wiser, freer, and more capable beings.

Yet such possibilities may easily be eclipsed if we remain fixated on building agentic systems. Concretely, non-agentic AIs can help humans in many ways (Cotton-Barratt and Douglas, 2024):

- generate a plan based on a context and goal given by a user (e.g. Google Maps);
- execute a plan or vision generated by a user (e.g. generate an image);
- synthesize information relevant to a user’s question (e.g. a medical concern);
- convert high-level instructions into actions (e.g. an email assistant);
- critique or improve a plan or document created by a user.

Such systems can take the form of “human-in-the-loop” systems where cognition is coordinated entirely by human preferences via sparse-but-robust intervention points (Kees and Janus, 2023), oracles that package reasoning for question-answering (Armstrong et al., 2012), AI Scientists that apply expert reasoning to domain-specific innovation (Bengio et al., 2025), dynamic contracts enabled by powerful models (Hoffman and Beato, 2025), decision support systems that help humans navigate complex choices, or ambient cognitive automation that improves the quality of perception and deliberation available to human cognition.

8 How Can We Steer AI Towards Human Amplification?

The endeavor to reorient AI development towards human amplification can begin from several angles. This section outlines a progression of undertakings, commencing with direct implementation and moving towards foundational research.

8.1 Constructing Agency-Increasing AI Systems

The most direct path to human amplification is to construct AI systems with this explicit purpose. Such development initiates a positive feedback loop – recursive human amplification – where augmented human capabilities enable the creation of progressively more effective support systems. Examples of such systems include:

- **AI Tutor:** interactive systems designed to provide personalized guidance, feedback, and support across a variety of domains, with the goal of enhancing human capability, decision-making, and well-being (Kim and Kim, 2020; Baillifard et al., 2025).
- **Cognitive Augmentation Tools:** systems that expand individual human cognitive functions (Kees and Janus, 2023), including creativity and problem solving, e.g. facilitating exploration

¹⁴Indeed, form factors may matter for adoption dynamics even when systems achieve functional equivalence. While a superintelligent AI tutor (aligned in the image of good human teachers) and a maximally adaptive textbook could achieve functional equivalence at their respective developmental asymptotes – both optimally scaffolding learning for individual students – their pathways of distributed adoption exhibit markedly different patterns of integration with human epistemic practices. The latter favors more organic assimilation into existing pedagogical ecologies. This distinction gestures toward a broader point: even within co-synthetic approaches, the ontological form factor of artificial systems shapes the sociology of their adoption in ways not captured by purely functional analysis.

of novel conceptual spaces; selecting and distilling complex information; making humans less susceptible to scams and misinformation (Buterin, 2023).

- **Ambient Intelligence:** automated systems designed to manage routine, logistical, or informational tasks with minimal user invocation, operating unobtrusively within a user’s environment or workflows (e.g., automated calendar scheduling) (Gams et al., 2019; Nahar and Kachnowski, 2023). Their contribution to human agency lies in the capacity to reduce cognitive load.
- **Collective Agency Platforms:** AI systems that bolster the capacities of groups (Galesic et al., 2023; Cui and Yasserli, 2024). This includes platforms to improve democratic deliberation by structuring discussions or synthesizing diverse viewpoints, and technologies for structured privacy which enable secure information sharing and collaborative work while preserving necessary confidentiality.

8.2 Benchmarks and Optimization Metrics

To promote community-wide efforts towards AI systems that increase human agency, the development of robust benchmarks and precise metrics is essential. Such evaluative frameworks serve to channel research endeavor – indeed, the very pursuit of defining and operationalizing agency for measurement stimulates further investigation into its constituent elements – and permit consistent tracking of advancements (Lambert, 2025).

HumanAgencyBench (Sturgeon et al., 2025) is a first version of such a benchmark, measuring how well the AI is at respecting the user’s autonomy. Sharma et al. (2026) document patterns of disempowerment in LLM usage. More ambitious benchmarks could include a human-in-the-loop, and measure how well the AI system helps the human achieve things that they care about (Laidlaw et al., 2025). This would incentivize the AI to extend human agency rather than replacing it. In contrast, most current evaluations measure how well an AI system is at doing a task by itself – optimizing this pushes us more towards autonomous agents (e.g. (Liu et al., 2023)).

Another set of evaluations could probe the AI’s capacity to develop and maintain an accurate model of the user’s knowledge, preferences, and intentions – a form of computational theory of mind – and its ability to reflect this understanding without over-determination or narrative imposition (Kosinski, 2024). Further, benchmarks should scrutinize the interaction itself, evaluating the AI’s proficiency in posing questions that promote user reflection and independent problem solving, or its capacity to help users uncover their own revealed preferences through interaction data (Qadri et al., 2024).

8.3 A deeper understanding of agency

While initial designs of agency-increasing AI can leverage existing conceptions of human volition and capability, long-term success requires a more detailed model of human agency itself. This calls for a concerted interdisciplinary effort, integrating insights from psychology on motivation (Brown, 2007) and self-determination (Deci and Ryan, 2004), neuroscience on the underpinnings of decision-making and action (Fellows, 2004; Haggard, 2008), and philosophy on the nature of autonomy and intentionality (Ashton, 2023).

A central question is how to distinguish AI assistance that genuinely expands a user’s capacity to act, from support that inadvertently leads to passivity, skill atrophy, or a diminished sense of ownership over outcomes. Delegating writing may boost productivity today while eroding critical thinking tomorrow; an AI that answers every question may leave one less able to formulate good questions in the first place.

Closely related is the “extended mind” question (Clark and Chalmers, 1998; Barandiaran and Pérez-Verdugo, 2025; Chiriatti et al., 2025): under what conditions is an AI system integrated as part of an individual’s cognitive system, rather than (perceived as) an external source of capability? The answer likely shapes whether the human experiences genuine agency or mere dependence. Determining what features of a cognitive extension make it one or the other is a central challenge for co-synthetic research.

9 Conclusions

We have argued that the current focus on building autonomous artificial agents presents distinct challenges regarding control, power concentration, and human displacement. A better aim is the responsible amplification of human agency, which is both necessary and nearly sufficient for a future worth wanting. Such amplification can come from artificial agents, but a change in aim opens up a much broader design space of non-agentic possibilities.

Even a partial shift could prove transformative, thanks to the compounding returns of *recursive human amplification*: the more we enhance our ability to understand and to act, the better positioned we are – collectively – to steer future technologies toward further amplification.

Ultimately, AI is a cognitive Copernican revolution that shows that intelligence is a much broader phenomenon than what goes on in human or animal agents (Ball, 2022). To realize the full scope of the revolution, we must not let our preconceptions from biological intelligence limit our imagination for what (artificial) intelligence can be. Assuming that intelligence has to come in the form of an agent, is like assuming that if the Earth is not the center of the universe, then the Sun must be. Galaxies are impossible.

References

- Aguirre, A. (2025). Keep the future human: Why and how we should close the gates to AGI and superintelligence, and what we should build instead. *SSRN Electronic Journal*.
- Allen, C. and Varner, G. (2000). Prolegomena to any future artificial moral agent. In Anderson, M. and Anderson, S. L., editors, *Machine Ethics: Creating an Ethical Intelligent Agent*, pages 175–186. AAAI Press, Menlo Park, CA. Reprinted and discussed in later collections on machine ethics.
- Amodei, D. (2024). Machines of loving grace: How AI could transform the world for the better. Blog essay, <https://darioamodei.com/machines-of-loving-grace>. October 2024.
- Anderson, M. and Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–26.
- Arendt, H. (1958). *The human condition*. University of Chicago press.
- Armstrong, S., Sandberg, A., and Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4):299–324.
- Asaro, P. M. (2008). From mechanisms of adaptation to intelligence amplifiers: The philosophy of w. ross ashby. In Husbands, P., Holland, O., and Wheeler, M., editors, *The Mechanical Mind in History*, pages 149–184. MIT Press, Cambridge, MA.
- Ashby, W. R. (1956). Design for an intelligence-amplifier. In Shannon, C. E. and McCarthy, J., editors, *Automata Studies*, volume 34 of *Annals of Mathematics Studies*, pages 215–233. Princeton University Press, Princeton, NJ.
- Ashton, H. (2023). Definitions of intent suitable for algorithms. *Artificial Intelligence and Law*, 31(3):515–546.
- Baillifard, A., Gabella, M., Lavenex, P. B., and Martarelli, C. S. (2025). Effective learning with a personal AI tutor: A case study. *Education and Information Technologies*, 30(1):297–312.
- Ball, P. (2022). *The Book of Minds: How to Understand Ourselves and Other Beings, from Animals to AI to Aliens*. University of Chicago Press.
- Barandiaran, X. E. and Pérez-Verdugo, M. (2025). Generative midtended cognition and artificial intelligence: Thinging with thinging things. *Synthese*, 205(4):1–24.
- BCG (2024). AI Agents: What They Are and Their Business Impact. Insight report, Boston Consulting Group. Accessed 7 December 2025.

- Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., et al. (2025). Superintelligent agents pose catastrophic risks: Can scientist AI offer a safer path? *arXiv preprint arXiv:2502.15657*.
- Bent, B. (2025). The term ‘agent’ has been diluted beyond utility and requires redefinition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '25)*, volume 8, pages 403–413. AAAI Press.
- Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind*. Cambridge University Press.
- Blili-Hamelin, B., Graziul, C., Hancox-Li, L., Hazan, H., El-Mhamdi, E.-M., Ghosh, A., Heller, K., Metcalf, J., Murai, F., Salvaggio, E., et al. (2025). Stop treating AGI’ as the north-star goal of AI research. *arXiv preprint arXiv:2502.03689*.
- Blili-Hamelin, B., Hancox-Li, L., and Smart, A. (2024). Unsocial intelligence: An investigation of the assumptions of AGI discourse. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 141–155.
- Bostrom, N. (2003). When machines outsmart humans. *Futures*, 35(7):759–764.
- Bostrom, N. (2012). What happens when our computers get smarter than we are? TED Talk, https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are. Accessed: 2025-05-16.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford.
- Bostrom, N. (2024). *Deep utopia: life and meaning in a solved world*. Ideapress Publishing.
- Brown, L. V. (2007). *Psychology of motivation*. Nova Publishers.
- Brynjolfsson, E. (2022). The Turing trap: The promise & peril of human-like artificial intelligence. *Daedalus*, 151(2):272–287.
- Bullock, J. B., Hammond, S., and Krier, S. (2025). AGI, governments, and free societies. *arXiv preprint arXiv:2503.05710*.
- Buterin, V. (2023). My techno-optimism. Web page.
- Carlsmith, J. (2022). Is power-seeking AI an existential risk? *arXiv preprint arXiv:2206.13353*.
- Carranza, A., Pai, D., Schaeffer, R., Tandon, A., and Koyejo, S. (2023). Deceptive alignment monitoring. *arXiv preprint arXiv:2307.10569*.
- Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., et al. (2023). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666.
- Chang, R. (2017). Hard choices. *Journal of the American Philosophical Association*, 3(1):1–21.
- Chiriatti, M., Bergamaschi Ganapini, M., Panai, E., Wiederhold, B. K., and Riva, G. (2025). System 0: Transforming artificial intelligence into a cognitive extension. *arXiv preprint arXiv:2506.14376*.
- Christiano, P., Shlegeris, B., and Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.
- Christman, J. (2009). *The politics of persons: Individual autonomy and socio-historical selves*. Cambridge University Press.
- Clark, A. (2025). Extending minds with generative AI. *Nature Communications*, 16(1):4627.
- Clark, A. and Chalmers, D. (1998). The extended mind. *analysis*, 58(1):7–19.
- Cockshott, W. P. and Cottrell, A. F. (1993). *Towards a New Socialism*. Spokesman Books, Nottingham.

- Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 3(4):1341–1350.
- Coeckelbergh, M. (2025). LLMs, truth, and democracy: An overview of risks. *Science and Engineering Ethics*, 31(1):1–13.
- Cohen, M. K., Hutter, M., and Osborne, M. A. (2022). Advanced artificial agents intervene in the provision of reward. *AI Magazine*, 43(3):282–293.
- Cotton-Barratt, O. and Douglas, R. (2024). Decomposing agency — capabilities without desires. LessWrong post, <https://www.lesswrong.com/posts/jpGHShgevmmTqXHy5/decomposing-agency-capabilities-without-desires>. Accessed: 2025-05-16.
- Crisp, R. (2021). Well-Being. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.
- Cui, H. and Yasserli, T. (2024). AI-enhanced collective intelligence. *Patterns*, 5(11).
- Dancy, J. (2000). *Practical Reality*. Oxford University Press, Oxford.
- Davis, R., Shrobe, H., and Szolovits, P. (1993). What is a knowledge representation? *AI Magazine*, 14(1).
- Deci, E. L. and Ryan, R. M. (2004). *Handbook of self-determination research*. University Rochester Press.
- Dewey, J. (1927). *The public and its problems*. Holt Publishers.
- Di Langosco, L. L., Koch, J., Sharkey, L. D., Pfau, J., and Krueger, D. (2022). Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR.
- Drago, L. and Laine, R. (2025). The intelligence curse. <https://intelligence-curse.ai/>. Accessed: 2025-05-16.
- Du, Y., Tiomkin, S., Kiciman, E., Polani, D., Abbeel, P., and Dragan, A. (2020). Ave: Assistance via empowerment. *Advances in Neural Information Processing Systems*, 33:4560–4571.
- Dung, L. (2025). Understanding artificial agency. *The Philosophical Quarterly*, 75(2):450–472.
- Ellis, E., Myers, V., Tuyls, J., Levine, S., Dragan, A., and Eysenbach, B. (2025). Training LLM agents to empower humans. *arXiv preprint arXiv:2510.13709*.
- Engelbart, D. C. (1962). Augmenting human intellect: A conceptual framework. Technical Report AFOSR-3233, Stanford Research Institute, Menlo Park, CA. Summary report prepared for the Air Force Office of Scientific Research.
- Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467.
- Feigenbaum, E. A. (1992). A personal view of expert systems: Looking back and looking ahead. *Expert Systems with Applications*, 5(3–4):193–201.
- Fellows, L. K. (2004). The cognitive neuroscience of human decision making: a review and conceptual framework. *Behavioral and cognitive neuroscience reviews*, 3(3):159–172.
- Figà-Talamanca, G. (2026). From AI to octopi and back: AI systems as responsive and contested scaffolds. In Müller, V. C., Dung, L., Löhr, G., and Rumana, A., editors, *Philosophy of Artificial Intelligence: The State of the Art*, volume 533 of *Synthese Library*. Springer Nature Switzerland. Forthcoming.
- Fischer, J. M. and Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge Studies in Philosophy and Law. Cambridge University Press.

- Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3):349–379.
- Foster, J. G. (2023). From thin to thick: Toward a politics of human-compatible AI. *Public Culture*, 35(3):417–430.
- Franklin, S. and Graesser, A. (1996). Is it an agent, or just a program? a taxonomy for autonomous agents. In Müller, J. P., Wooldridge, M. J., and Jennings, N. R., editors, *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages (ATAL-96)*, volume 1193 of *Lecture Notes in Computer Science*, pages 21–35, Berlin, Heidelberg. Springer.
- Freiesleben, T. and Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109.
- Friston, K. J., Daunizeau, J., Kilner, J., and Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biological Cybernetics*, 102(3):227–260.
- Galesic, M., Barkoczi, D., Berdahl, A. M., Biro, D., Carbone, G., Giannoccaro, I., Goldstone, R. L., Gonzalez, C., Kandler, A., Kao, A. B., et al. (2023). Beyond collective intelligence: Collective adaptation. *Journal of the Royal Society interface*, 20(200):20220736.
- Gams, M., Gu, I. Y.-H., Härmä, A., Muñoz, A., and Tam, V. (2019). Artificial intelligence and ambient intelligence. *Journal of Ambient Intelligence and Smart Environments*, 11(1):71–86.
- Goertzel, B. (2001). *Creating internet intelligence: Wild computing, distributed digital consciousness, and the emerging global brain*, volume 18. Springer Science & Business Media.
- Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers*, volume 6, pages 31–88. Elsevier.
- Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, 9(12):934–946.
- Harari, Y. N. (2024). AI will take over human systems from within. *Noema Magazine*. Interview by Nathan Gardels.
- Hardt, M. and Mandler-Dünner, C. (2023). Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*.
- Heins, C., Van de Maele, T., Tschantz, A., Linander, H., Markovic, D., Salvatori, T., Pezzato, C., Çatal, O., Wei, R., Koudahl, M. T., Perin, M., Friston, K. J., Verbelen, T., and Buckley, C. L. (2025). Axiom: Learning to play games in minutes with expanding object-centric models. *arXiv preprint arXiv:2505.24784*.
- Heylighen, F. (2011). Conceptions of a global brain: an historical review. In Grinin, L. E., Carneiro, R. L., Korotayev, A. V., and Spier, F., editors, *Evolution: Cosmic, Biological, and Social*, pages 274–289. Uchitel Publishing, Volgograd.
- Heylighen, F. and Lenartowicz, M. (2017). The global brain as a model of the future information society: An introduction to the special issue. *Technological Forecasting and Social Change*, 114:1–6.
- Hoffman, R. and Beato, G. (2025). *Superagency: What Could Possibly Go Right with Our AI Future*. Authors Equity.
- Holmes, A. (2025). The seven kinds of AI agents. *The Information*. Online article, accessed 7 December 2025.
- Hou, Y., Xiong, D., Jiang, T., Song, L., and Wang, Q. (2019). Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of psychosocial research on cyberspace*, 13(1).
- Hsu, Y.-C., Huang, T.-H. K., Verma, H., Mauri, A., Nourbakhsh, I., and Bozzon, A. (2022). Empowering local communities using artificial intelligence. *Patterns*, 3(3):100449.

- Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media.
- Irving, G., Christiano, P., and Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.
- Jin, W., Vincent, N., and Hamarneh, G. (2025). AI for just work: Constructing diverse imaginations of AI beyond "replacing humans". *arXiv preprint arXiv:2503.08720*.
- Joy, B. (2000). Why the future doesn't need us. *Wired*, 8(4). Available at <https://www.wired.com/2000/04/joy-2/>.
- Kapoor, S., Kolt, N., and Lazar, S. (2025). Build agent advocates, not platform agents. *arXiv preprint arXiv:2505.04345*.
- Kasirzadeh, A. (2025). Two types of AI existential risk: decisive and accumulative. *Philosophical Studies*, pages 1–29.
- Kasirzadeh, A. and Gabriel, I. (2025). Characterizing AI agents for alignment and governance. *arXiv preprint arXiv:2504.21848*.
- Kees, N. and Janus (2023). Cyborgism. LessWrong post, <https://www.lesswrong.com/posts/bxt7uCiHam4QXrQAA/cyborgism>. Accessed: 2025-05-16.
- Kim, W.-H. and Kim, J.-H. (2020). Individualized AI tutor based on developmental learning networks. *IEEE Access*, 8:27927–27937.
- Koralus, P. (2025). The philosophic turn for AI agents: Replacing centralized digital rhetoric with decentralized truth-seeking. *arXiv preprint arXiv:2504.18601*.
- Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. OUP Oxford.
- Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.
- Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J., Brna, A. P., Chakravarthi Raja, S., Cheney, N., Clune, J., et al. (2022). Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210.
- Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. (2025). Gradual disempowerment: Systemic existential risks from incremental AI development. *arXiv preprint arXiv:2501.16946*.
- Kurzweil, R., Richter, R., Kurzweil, R., and Schneider, M. L. (1990). *The age of intelligent machines*, volume 579. MIT press Cambridge.
- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Von Arx, S., et al. (2025). Measuring AI ability to complete long tasks. *arXiv preprint arXiv:2503.14499*.
- Laidlaw, C., Bronstein, E., Guo, T., Feng, D., Berglund, L., Svegliato, J., Russell, S., and Dragan, A. (2025). Assistancezero: Scalably solving assistance games. *arXiv preprint arXiv:2504.07091*.
- Lambert, K. G. (2006). Rising rates of depression in today's society: consideration of the roles of effort-based rewards and enhanced resilience in day-to-day functioning. *Neuroscience & Biobehavioral Reviews*, 30(4):497–510.
- Lambert, N. (2025). Brakes on an intelligence explosion. [Interconnects.ai](https://interconnects.ai).
- Last, C. (2020). Global brain: Foundations of a distributed singularity. In Korotayev, A. V. and LePoire, D., editors, *The 21st Century Singularity and Global Futures: A Big History Perspective*, pages 363–375. Springer, Cham.
- LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

- Legg, S. and Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17:391–444.
- Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.
- Lem, S. (1964). Intellectronics. In *Summa Technologiae*, pages 137–178. University of Minnesota Press, Minneapolis, MN. First published in Polish in 1964; translated by Joanna Zylińska.
- Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
- Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1):4–11.
- Licklider, J. C. R. (1965). *Libraries of the Future*. The M.I.T. Press, Cambridge, Massachusetts. Final report of a two-year research project sponsored by the Council on Library Resources.
- Lighthill, J. (1973). Artificial intelligence: A general survey. In *Artificial Intelligence: A Paper Symposium*, pages 1–21. Science Research Council, London.
- List, C. and Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.
- Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al. (2023). Agentbench: Evaluating LLMs as agents. *arXiv preprint arXiv:2308.03688*.
- London, A. J. and Heidari, H. (2024). Beneficent intelligence: a capability approach to modeling benefit, assistance, and associated moral failures through AI systems. *Minds and Machines*, 34(4):41.
- MacAskill, M., Bykvist, K., and Ord, T. (2020). *Moral uncertainty*. Oxford University Press.
- Macmillan-Scott, O. and Musolesi, M. (2025). (ir)rationality in AI: state of the art, research challenges and open questions. *Artificial Intelligence Review*, 58(11):352.
- Maes, P., editor (1990). *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. MIT Press, Cambridge, MA.
- Maes, P. (2017). Augmenting the human experience. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '17, page 1, New York, NY, USA. Association for Computing Machinery.
- McAdams, D. P. (2001). The psychology of life stories. *Review of general psychology*, 5(2):100–122.
- McCarthy, J. (1973). Review of “artificial intelligence: A general survey”. <https://www-formal.stanford.edu/jmc/reviews/lighthill/lighthill.html>. Accessed: 2025-05-19.
- McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Technical report, Dartmouth College. Reprinted in *AI Magazine*, 27(4):12–14, 2006.
- McShane, M., Nirenburg, S., and English, J. (2024). *Agents in the Long Game of AI: Computational Cognitive Modeling for Trustworthy, Hybrid AI*. MIT Press, Cambridge, MA. Open-access MIT Press monograph.
- Mill, J. S. (1859). *On liberty*. John W. Parker and Son.
- Mill, J. S. (1865). *Considerations on representative government by John Stuart Mill*. Longman, Green, Longman, Roberts, and Green.
- Mitchell, M., Ghosh, A., Luccioni, A. S., and Pistilli, G. (2025). Fully autonomous AI agents should not be developed. *arXiv preprint arXiv:2502.02649*.
- Mitelut, C., Smith, B., and Vamplew, P. (2023). Intent-aligned AI systems deplete human agency: the need for agency foundations research in AI safety. *arXiv preprint arXiv:2305.19223*.

- Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Penguin.
- Nahar, J. K. and Kachnowski, S. (2023). Current and potential applications of ambient artificial intelligence. *Mayo Clinic Proceedings: Digital Health*, 1(3):241–246.
- Nelson, A. (2023). Thick alignment. Keynote address at the 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).
- Newell, A. (1963). *A guide to the general problem-solver program gps-2-2*. Rand Corporation.
- Newell, A. and Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ. Prentice-Hall Series in Automatic Computation.
- Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126.
- Nielsen, M. and Matuschak, A. (2019). How can we develop transformative tools for thought? <https://numinous productions/ttft/>. Online essay.
- Nilsson, N. J. (1971). *Problem-Solving Methods in Artificial Intelligence*. McGraw-Hill, New York. McGraw-Hill Computer Science Series.
- Nilsson, N. J. (2005). Human-level artificial intelligence? be serious! *AI Magazine*, 26(4):71–83.
- Noller, J. (2024). Extended human agency: towards a teleological account of AI. *Humanities and Social Sciences Communications*, 11(1):1–7.
- Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Harvard University Press.
- Omohundro, S. M. (2018). The basic AI drives. In *Artificial intelligence safety and security*, pages 47–55. Chapman and Hall/CRC.
- Parkes, D. C. and Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272. Introduces the notion of *machina economicus*.
- Parson, E. A. (2020). Max – a thought experiment: Could AI run the economy better than markets? UCLA School of Law, Law-Econ Research Paper Law-Econ Research Paper No. 20-02, UCLA School of Law. Also available via AI Pulse.
- Pateman, C. (1975). *Participation and democratic theory*. Cambridge University Press.
- Prunkl, C. (2022). Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence*, 4(2):99–101.
- Qadri, R., Mirowski, P., Gabriellan, A., Mehr, F., Gupta, H., Karimi, P., and Denton, R. (2024). Dialogue with the machine and dialogue with the art world: Evaluating generative AI for culturally-situated creativity. *arXiv preprint arXiv:2412.14077*.
- Rao, A. S. and Georgeff, M. P. (1995). BDI agents: From theory to practice. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, pages 312–319, San Francisco, CA. MIT Press.
- Raz, J. (1999). Explaining normativity: On rationality and the justification of reason. *Ratio*, 12(4):354–379.
- Reinhardt, B. (2025). Humanoid robots in manufacturing. SpecTech Blog.
- Rheingold, H. (2000). *Tools for Thought: The History and Future of Mind-Expanding Technology*. MIT Press, Cambridge, MA. Revised edition.
- Robeyns, I. and Byskov, M. F. (2020). The capability approach. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition.
- Russell, S. (2016). Rationality and intelligence: A brief update. In Müller, V. C., editor, *Fundamental Issues of Artificial Intelligence*, volume 376 of *Synthese Library*, pages 9–28. Springer, Cham.

- Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson.
- Russell, S. J. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall.
- Ryan, R. M. and Deci, E. L. (2024). Self-determination theory. In *Encyclopedia of quality of life and well-being research*, pages 6229–6235. Springer.
- Sartre, J.-P. (1948). *Existentialism is a Humanism*. Methuen and Co.
- Scanlon, T. M. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press, Cambridge, MA.
- Sen, A. (1985). *Commodities and Capabilities*. North-Holland, Amsterdam.
- Sen, A. (1993). Capability and well-being. *The quality of life*, 30(1):270–293.
- Setiya, K. (2007). *Reasons without Rationalism*. Princeton University Press, Princeton, NJ.
- Shah, R. (2018). Coherence arguments do not entail goal-directed behavior. AI Alignment Forum. Part of the Value Learning sequence.
- Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. (2022). Goal misgeneralization: Why correct specifications aren’t enough for correct goals. *arXiv preprint arXiv:2210.01790*.
- Shao, Y., Zope, H., Jiang, Y., Pei, J., Nguyen, D., Brynjolfsson, E., and Yang, D. (2025). Future of work with AI agents: Auditing automation and augmentation potential across the U.S. workforce. *arXiv preprint arXiv:2506.06576*.
- Sharkey, L. (2022). Circumventing interpretability: How to defeat mind-readers. *arXiv preprint arXiv:2212.11415*.
- Sharma, M., McCain, M., Douglas, R., and Duvenaud, D. (2026). Who’s in charge? disempowerment patterns in real-world llm usage. *arXiv preprint arXiv:2601.19062*.
- Simon, H. A. (1965). *The Shape of Automation for Men and Management*. Harper & Row, New York. Based on the Katz Lectures delivered at the University of Pittsburgh, 1963.
- Simon, H. A. (1983). Why should machines learn? In *Machine learning*, pages 25–37. Elsevier.
- Slooman, A. (1993). The mind as a control system. *Royal Institute of Philosophy Supplement*, 34:69–110.
- Soares, N. (2023). Would we even want AI to solve all our problems? LessWrong post.
- Soares, N. and Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. MIRI Technical Report 2014-8, Machine Intelligence Research Institute. Original research agenda draft; later revised and retitled.
- Sohl-Dickstein, J. (2023). The hot mess theory of AI misalignment: More intelligent agents behave less coherently. <https://sohl-dickstein.github.io/2023/03/09/coherence.html>. Blog post.
- Solomonoff, R. J. et al. (2006). Machine learning-past and future. *Dartmouth, NH, July*.
- Sparks, J. and Wright, A. T. (2025). Models of rational agency in human-centered AI: the realist and constructivist alternatives. *AI and Ethics*, 5(3):3321–3328.
- Stanley, K. O. and Lehman, J. (2015). *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer, Cham.
- Strange, A. and da Costa, J. (2024). Every white-collar role will have an AI copilot. then an AI agent. *Andreessen Horowitz (a16z) Future*. Online essay.
- Sturgeon, B., Hyams, L., Samuelson, D., Vorster, E., Haimes, J., and Anthis, J. R. (2025). Human-agencybench: Do language models support human agency? In *Workshop on Datasets and Evaluators of AI Safety*.

- Suchman, L. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge University Press.
- Sukharevsky, A., Kerr, D., Hjartar, K., Hämäläinen, L., Bout, S., Di Leo, V., and Dagorret, G. (2025). Seizing the agentic AI advantage: A ceo playbook to solve the gen AI paradox and unlock scalable impact with AI agents. Report, McKinsey & Company, QuantumBlack, AI by McKinsey. 28-page report.
- Sutton, R. S., Barto, A. G., et al. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134.
- Tong, R. J. (2026). From augmentation to symbiosis: A review of human-AI collaboration frameworks, performance, and perils. *arXiv preprint arXiv:2601.06030*.
- Trask, A., Bluemke, E., Collins, T., Drexler, B. G. E., Cuervas-Mons, C. G., Gabriel, I., Dafoe, A., and Isaac, W. (2020). Beyond privacy trade-offs with structured transparency. *arXiv preprint arXiv:2012.08347*.
- Tsai, L. L., Pentland, A. S., Braley, A., Chen, N. L., Enríquez, J. R., and Reuel, A. (2024). Generative AI for pro-democracy platforms. Impact paper, MIT GOV/LAB; MIT Exploration of Generative AI.
- Vaintrob, L. and Cotton-Barratt, O. (2025). AI tools for existential security. Technical report, Forethought Foundation.
- Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. In Bertolotti, C. M., editor, *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, pages 11–22, Cleveland, OH. NASA Lewis Research Center. NASA Conference Publication 10129.
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):233.
- Weyl, E. G., Tang, A., and the Plurality Community (2024). *Plurality: The Future of Collaborative Technology and Democracy*. Independently published. Open, git-based collaborative book project; see <https://www.plurality.net/>.
- White, B., Clark, A., Guènin-Carlut, A., Constant, A., and Di Paolo, L. D. (2025). Shifting boundaries, extended minds: Ambient technology and extended allostatic control. *Synthese*, 205(2):1–28.
- Wiener, N. (1954). *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin, Boston. Revised edition.
- Wooldridge, M. and Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152.
- Yager, K. G. (2024). Towards a science exocortex. *Digital Discovery*, 3(10):1933–1957.
- Zhi-Xuan, T., Carroll, M., Franklin, M., and Ashton, H. (2024). Beyond preferences in AI alignment. *Philosophical Studies*, pages 1–51.

A Frequently Asked Questions

A.1 Isn't solving alignment enough? Why do we need to shift the aims as well?

Some would argue that our critique of the iso-synthetic approach is based on a misunderstanding of what a world with powerful, fully aligned AI looks like. For example, Soares (2023) argues that a powerful (singleton) AI agent that is properly aligned with human values, will understand that humans need to do real things in order to be happy, and so will leave sufficient room for human agency. We agree that this could be a good outcome, but question whether building an aligned AI agent is the best way of getting there. There seems to be something backward with the idea of ceding all control to an AI agent only for it to cede some control back to us, especially given the technological and sociological problems discussed in Section 4.

A.2 Can non-agentic AI be competitive?

A common pushback is that humans will soon be outcompeted by artificial agents, making any attempt to keep humans relevant futile. This argument assumes (1) artificial agents will soon be much more competent than humans, and (2) humans will stay roughly the same. But these premises are unlikely to hold jointly: as discussed in Section 7, the same technological trajectory that could produce full AI agents will also yield powerful cognitive tools that amplify human capabilities (Reinhardt, 2025; Kwa et al., 2025).

Humans amplified by such tools can retain authentic agency, as long as they maintain high-level understanding and the responsibility to direct their actions. They may lack full insight into how the information they rely on was produced or how their commands are executed – but the same is true for current CEOs and government leaders, whom we readily credit with significant agency. The only disadvantage compared to purely artificial agents is that high-level decision-making resides with a human. But unless we are willing to give up on alignment altogether, this is inevitable: axiological uncertainty, self-creation, and the inherent value of human agency (Section 5) mean that no aligned artificial agent can do more than amplified humans can do for themselves.

A.3 Aren't good AI agents needed to defend against bad AI agents?

Bostrom (2014) argues that even if the risks from artificial agents are real, the alternatives are worse. "Good" AI agents are the only viable defense against "bad" AI agents created by rogue actors. The validity of this argument hinges on empirical questions like the advantages artificial agents give you over non-agentic AI systems and other mechanisms of control (like regulation). Is it possible to sufficiently monitor AI development to prevent rogue actors from creating artificial agents that threaten the power equilibrium, without relying on artificial agents oneself? Not necessarily. As argued above, agentic AI may offer only modest advantages over non-agentic AI. And if not, a co-synthetic approach is compatible with a focus on agents when they are the best tool for the job.

A.4 Is it necessarily bad to build AI agents?

No, for some use cases, agents may well be the most appropriate form factor. We are only saying that we shouldn't make agents the primary goal. We don't object to agents when they are the best tool for the job.

A.5 It's impossible to shift the direction of AI research

It's true that AI research is shaped by strong economic incentives, some of which push the field towards autonomous artificial agents. However, at least at the moment, the drive towards agents seems to come as much from research excitement as from economic calculation. Many economically valuable use cases do not require full agents. So if we can shift research excitement towards human agency amplification rather than autonomous agents, a partial shift would likely be possible.

A.6 What if the shift is just partial?

Even a partial shift in direction towards human agency amplification can spark a self-sustaining cycle of recursive human amplification, where increases in our understanding and our ability to steer technology lead to technology that increases our agency even more. Any such increase will enable us to better handle the challenges the world – and technological development – throw at us.