NICLAS BRAUN / PREPRINT

VERSION 1.1 / JULY 2026

THE END OF NECESSARY ADULTHOOD

A Theory of Human Agency Under Omnipresent Superintelligence

SCIENTIFIC PREPRINT

AI ethics · developmental theory · human agency

Conditional analysis · no empirical ASI claim

NICLAS BRAUN · INDEPENDENT RESEARCHER

Contents

A conditional theory of responsibility displacement, developmental atrophy, and guardianship equilibrium.

Scope and disclosure statement

This is a conceptual and prospective article. Artificial superintelligence as defined here does not presently exist, and the paper reports no direct empirical observation of a superintelligent society. Its claims are therefore conditional: the analysis asks what may follow **if** highly capable, reliable, ubiquitous, and action-capable artificial systems become the default infrastructure through which human needs are anticipated, decisions are optimized, and adverse consequences are prevented. Evidence from developmental psychology, life-course sociology, human–automation research, cognitive offloading, and the philosophy of agency is used to identify plausible mechanisms and testable precursor effects. It is not treated as proof of the proposed long-run outcome.

The terms adulthood, childhood, and infantilization are used in a functional and institutional sense, not as diagnoses, insults, or claims that one culture's life course is universal. The paper does not argue that hardship, deprivation, or preventable danger is developmentally desirable. Its concern is whether societies can remove involuntary burdens without also removing the conditions under which people learn to author consequential choices, answer for outcomes, and govern common life.

Abstract

Debates about advanced artificial intelligence commonly ask whether increasingly capable systems will remain aligned with human values and subject to human control. This article examines the reciprocal governance problem: whether humans would retain the capacities required for meaningful agency if artificial systems became more competent, reliable, and pervasive than human decision-makers. It develops the **end of necessary adulthood hypothesis** as a conditional theory, not a forecast. Under the stipulated boundary condition of omnipresent artificial superintelligence (ASI), functional adulthood may cease to be a generally necessary social achievement because practical judgment, foresight, coordination, and consequence management can be delegated to a persistently superior system.

The article defines functional adulthood not by age, independence, or legal status, but by a socially cultivated bundle of capacities: consequential authorship, judgment under uncertainty, care, answerability, self-regulation, and collective agency. Using a purposive synthesis of life-course research, theories of agency and self-determination, cognitive offloading, human–automation research, and scholarship on moral responsibility, it specifies a five-part mechanism: concentration of actionable competence; rational deference; displacement of responsibility-bearing practice; intergenerational attenuation; and institutional lock-in. These mechanisms could generate a **guardianship equilibrium** in which people retain formal rights and high welfare while losing the practiced competence and recognized authority needed to govern consequential processes without machine mediation.

The framework distinguishes three rival trajectories—infantilization, liberation, and bifurcation—and derives seven empirically discriminating propositions. It does not infer long-run decline from present-day AI studies; existing evidence is used only to establish plausible component mechanisms and research designs. The central implication is that alignment with human preferences and preservation of human agency are related but non-equivalent objectives. Agency-preserving governance should protect bounded domains in which human judgment remains causally effective, answerable, and open to contestation, while avoiding the romanticization of hardship or unaided action.

Keywords: artificial superintelligence; functional adulthood; human agency; responsibility; automation; deskilling; meaningful human control; AI governance

1. Introduction

A civilization can eliminate danger without preserving agency. It can solve problems so successfully that the human capacities once needed to confront them are no longer cultivated. This paper examines the limiting case of that possibility.

Most debates about advanced artificial intelligence are organized around the behavior of the machine. Will it pursue the goals humans intend? Will it remain controllable? Will it deceive, manipulate, accumulate power, or produce catastrophic error? These questions are indispensable. They nevertheless leave a second-order question underdeveloped:

> If a superior intelligence reliably carries the burdens through which human beings learn judgment, responsibility, and self-government, what remains of adulthood?

The question is not whether adults will become less intelligent in a psychometric sense, nor whether every act of delegation is corrosive. Human civilization is built on cognitive and practical extension. Writing, maps, institutions, markets, scientific instruments, and computers allow people to achieve more than unaided individuals could accomplish. External support can expand agency rather than diminish it [1,2]. The relevant concern begins where extension becomes comprehensive substitution: not when a tool helps a person to act, but when an infrastructure persistently anticipates the need to act, selects the action, performs it, and neutralizes its consequences.

This boundary case matters because adulthood is not produced by birthdays alone. In life-course research, entry into adulthood varies historically and culturally, and is associated with transitions in responsibility, independent decision-making, work, care, and social role [3–5]. In theories of agency, human beings are not merely recipients of outcomes; they exercise intentionality, forethought, self-regulation, and reflection, both individually and collectively [6]. Self-determination theory similarly treats autonomy and competence as basic psychological needs rather than optional decorative features of well-being [7,8]. These traditions differ in method and normative commitments, but they converge on a minimal observation: mature agency is exercised and sustained through structured participation in consequential life.

An omnipresent ASI could alter that structure. Assume a system that is better than nearly every human at forecasting, planning, diagnosis, negotiation, instruction, resource allocation, and conflict prevention; is available at negligible marginal cost; can act through digital and physical infrastructure; is highly reliable; and is trusted by institutions. In such a world, following the system's recommendation would usually be prudent. Allowing a fallible person to overrule it could appear irresponsible. Parents, employers, insurers, hospitals, courts, and governments would face pressure to adopt the system precisely because it reduces preventable harm. The stronger its performance, the more compelling the pressure.

The resulting risk is not straightforward coercion. It is a **responsibility displacement loop**. Superior performance makes deference rational. Repeated deference reduces opportunities to acquire and demonstrate independent competence. Reduced competence makes future deference more rational. Institutions then redesign themselves around the assumption that the system will remain present. Across generations, the exception becomes the norm: human judgment is retained ceremonially, while practical authority migrates to the guardian.

This article calls the endpoint a **guardianship equilibrium**. It is an equilibrium because no individual actor has a strong incentive to leave it. A person who refuses assistance bears avoidable costs and risks. An institution that permits unassisted discretion may be judged negligent. A government that forgoes superior prediction may appear reckless. Yet the aggregate result can be

a civilization in which humans retain formal rights without retaining the practiced capacities required to exercise them.

The argument is conditional, not predictive. Existing generative AI is not ASI, and contemporary studies of automation cannot establish the psychology of a hypothetical solved world. The value of the analysis lies elsewhere: it identifies a neglected failure mode, separates it from adjacent concerns, derives observable precursor effects, and proposes design conditions under which increasingly capable systems might augment human agency rather than make it obsolete.

1.1 Research questions

The paper addresses four questions:

1. What constitutes functional adulthood when biological age and legal majority are set aside?
2. Through which mechanisms could pervasive machine competence reduce the development and maintenance of adult agency?
3. Under what conditions would delegation produce infantilization, liberation, or a bifurcation between governing and governed populations?
4. Which institutional and design principles could preserve human self-government without preserving unnecessary suffering, error, or labor?

1.2 Contribution

The article makes five contributions. First, it defines functional adulthood as a socially reproduced, responsibility-bearing capacity rather than an age category. Second, it distinguishes cognitive offloading, attributional shifts in responsibility, and the structural displacement of responsibility-bearing practice. Third, it develops the end of necessary adulthood hypothesis and the concept of a guardianship equilibrium. Fourth, it specifies rival trajectories and seven discriminating propositions. Fifth, it identifies preservation of human authorship as an analytically distinct objective within AI safety and governance.

The nearest adjacent accounts treat AI deskilling as a structural property of environments that cultivate or suppress particular capacities. The present framework changes the object and scale of analysis. Its object is not a single occupational, cognitive, or moral skill but a bundle of developmental and political capacities associated with functional adulthood. Its boundary condition is not present-day task automation but a society organized around persistently superior, action-capable machine intelligence. Its proposed mechanisms include intergenerational transmission and institutional lock-in, and its outcomes include both a population-level guardianship equilibrium and stratification in the distribution of decision authority. These extensions are hypotheses to be tested, not established consequences of current AI.

1.3 Method

This is a theory-building conceptual article, not a systematic review, meta-analysis, or empirical study. Sources were selected purposively because they bear on a component of the proposed causal account: developmental and life-course formation; agency and self-determination; cognitive offloading and extended cognition; human factors and automation; structural deskilling; and responsibility in autonomous systems. The selection is mechanism-oriented and does not claim exhaustive coverage of any one literature.

The analysis proceeds through six operations: definition of the focal construct; specification of the technological boundary condition; separation of adjacent mechanisms; construction of a causal sequence; comparison of rival trajectories; and derivation of observable implications and disconfirmation criteria. The cited literatures provide evidence about components of the proposed mechanism, not direct evidence for its ASI-scale conclusion. The article therefore separates three epistemic levels throughout:

- **Established adjacent findings:** evidence about human agency, development, offloading, automation, and responsibility in existing settings.
- **Mechanistic extrapolations:** reasoned claims about how these effects may interact as system competence, ubiquity, and institutional reliance increase.
- **Boundary-case hypotheses:** claims about a society organized around omnipresent ASI, which remain empirically unverified.

No probability is assigned to the stipulated future and no parameter is estimated from present-day evidence. The framework is intended to support later operationalization, comparison, and attempted falsification.

2. The boundary condition: what an omnipresent, “solving” ASI means

2.1 A stipulated analytical limit, not a forecast

The phrase an ASI that has solved all questions is rhetorically vivid but scientifically imprecise. Some questions are not instrumental problems with uniquely correct answers. Political priorities, moral commitments, aesthetic judgments, and conceptions of a good life can remain plural even under perfect factual knowledge. The scenario examined here is therefore narrower.

Omnipresent ASI denotes an artificial system, or interoperating ecology of systems, that satisfies six conditions:

1. **Broad competence:** it exceeds the best available human performance across nearly all instrumental domains relevant to personal and institutional action.
2. **Predictive depth:** it models likely consequences, individual preferences, and system interactions more accurately than ordinary human decision-makers.
3. **Ubiquity:** it is continuously accessible across education, work, health, finance, care, administration, and governance.
4. **Action capability:** it can not only advise but also coordinate and execute through networked institutions and physical systems.
5. **High reliability:** its error rate is sufficiently low that routine disagreement with it is difficult to justify on expected-outcome grounds.
6. **Protective backstop:** it can detect, correct, or contain many harmful consequences of human error before those consequences become irreversible.

The analysis does not require literal omniscience or metaphysical certainty. It requires a persistent competence gap large enough that institutional deference to ASI becomes the default rational policy.

2.2 A functionally solved world

A **functionally solved world** is one in which the burdens that historically forced most people to exercise practical judgment can usually be delegated. Food, shelter, diagnosis, scheduling, transportation, education, financial planning, conflict mediation, legal navigation, and public administration need not be perfect. They need only be sufficiently well managed that unaided human intervention is normally inferior and unnecessary.

Three limits remain.

First, normative disagreement may persist. ASI can map consequences and compatibility among values without deciding which values ought to bind everyone. Second, scarcity may persist for

positional or inherently rival goods. Third, catastrophe may remain possible. The scenario is not paradise; it is a world in which ASI is the dominant provider of instrumental competence and protection.

These limits strengthen rather than dissolve the research question. If humans retain nominal authority over values while losing the competence and practice required to deliberate, contest, and implement them, formal sovereignty may become hollow.

2.3 Assistance, delegation, substitution, and guardianship

The argument depends on distinctions that are often blurred.

Assistance occurs when a system expands the agent's capacity while the agent remains substantially engaged in framing, evaluating, and owning the act.

Delegation occurs when an agent authorizes another entity to perform a task while retaining the competence and authority to evaluate performance, revise the mandate, and resume control.

Substitution occurs when the system performs the activity and the person's corresponding capacity becomes unnecessary to the outcome.

Guardianship occurs when substitution extends beyond tasks to the management of the person's interests, risks, choices, and consequences. A guardian does not merely answer a question. It structures the environment so that the person rarely needs to confront the question independently.

These categories describe functions, not intentions. A benevolent system can become a guardian precisely because it protects the user effectively.

---

3. Adulthood as a socially reproduced capacity

3.1 Beyond biological and legal adulthood

Biological maturity and legal majority are administratively clear but analytically insufficient. A person can be legally adult while lacking control over major decisions, and a young person can exercise substantial responsibility before legal majority. Life-course scholarship treats adulthood as historically variable, socially structured, and shaped by the interaction between personal agency and institutional opportunity [3–5].

Arnett identifies accepting responsibility for oneself and making independent decisions among the important subjective criteria associated with becoming an adult in societies where emerging adulthood is recognized [3]. Shanahan emphasizes that pathways into adult roles vary across historical conditions and that young people actively shape biographies within structured opportunities and constraints [4]. Elder's life-course account similarly locates development in linked lives, historical time, and the interplay between agency and context [5]. These accounts do not establish one universal checklist. They do establish that adulthood is reproduced through social arrangements, not simply released by maturation.

This paper uses **functional adulthood** to denote a cluster of capacities through which a person participates as an author and bearer of consequence in personal and common life:

- **Consequential authorship:** forming or endorsing purposes that materially guide action.
- **Judgment under uncertainty:** deciding when evidence, values, and outcomes are incomplete or contested.

- **Provision and care:** taking durable responsibility for the needs of others and for shared institutions.
- **Answerability:** recognizing oneself as an appropriate target of reasons, explanation, credit, blame, and repair.
- **Self-regulation:** sustaining commitments through effort, delay, conflict, and changing circumstances.
- **Collective agency:** participating competently in the construction and revision of common rules.

No single episode must instantiate all six. The concern is whether a life contains a sufficiently rich ecology of activities through which they are learned, exercised, corrected, and socially recognized.

Functional adulthood does not mean self-sufficiency. Dependence, disability, assistive technology, care, and extensive social interdependence are compatible with full agency and adult standing. Relational accounts of autonomy emphasize that agency is socially constituted and can be enabled by supportive relationships rather than diminished by them [9]. The relevant distinction is therefore not between assisted and unassisted persons. It is between arrangements that expand a person's voice, recognized authority, contestability, and causal participation and arrangements that substitute for those capacities while continuing to treat the person as the nominal principal.

The inclusion of collective agency is deliberate. Functional adulthood is not exhausted by private competence; it also concerns appearing and acting with others in a shared public world, a dimension central to Arendt's account of action and political life [10].

3.2 Agency is exercised, not merely possessed

Bandura's account of human agency emphasizes intentionality, forethought, self-reactiveness, and self-reflectiveness, while distinguishing direct, proxy, and collective modes of agency [6]. Proxy agency—securing desired outcomes through another actor—is not inherently deficient. It is essential in complex societies. The problem arises when proxy agency displaces direct and collective agency so comprehensively that people can no longer evaluate the proxy or act without it.

Self-determination theory supplies a related distinction. Autonomy is not mere independence; it concerns volition and self-endorsement. Competence concerns effective interaction with the environment, while relatedness concerns meaningful connection with others [7,8]. An ASI could maximize comfort and preference satisfaction while weakening opportunities to experience competence or to form preferences through self-authored engagement. Welfare delivery and self-determination are therefore related but non-identical objectives.

The classic field experiment by Langer and Rodin is limited in population, era, and scale, yet remains conceptually instructive: nursing-home residents given greater choice and personal responsibility showed improvements in alertness, participation, and well-being relative to a comparison condition [11]. The result should not be generalized to ASI societies as a causal estimate. It illustrates a narrower point: decision-free environments can affect functioning even when the decisions removed are mundane and caretaking is benevolent.

3.3 The responsibility ecology

Adult capacities are rarely taught by explicit instruction alone. They develop within a **responsibility ecology**: recurring situations in which a person's choice matters, consequences are not wholly transferred elsewhere, other people can make claims on the person, and competence must be demonstrated over time.

For an episode to provide full responsibility-bearing practice, four properties are especially important:

1. **Authorship:** the person does more than ratify a conclusion already fixed by a superior authority.
2. **Stakes:** the outcome matters beyond the appearance of participation.

3 **Causal efficacy:** different choices can produce meaningfully different outcomes.

4 **Answerability:** the person is expected to explain, learn from, repair, or otherwise own the result.

This is a conjunctive concept, not a validated psychometric scale. A simulation can create difficulty without stakes. A formal vote can preserve appearance without causal efficacy. A choice architecture can preserve options without authorship if one option is effectively mandatory. Liability can impose answerability without control. Developmentally meaningful responsibility requires an adequate combination of all four.

The account is compatible with responsibility theories that connect moral responsibility to reasons-responsive guidance control rather than to bare causal involvement [12]. Its distinct purpose is developmental: to ask whether the relevant control capacities continue to be formed when superior systems routinely act in their place.

3.4 Necessary and voluntary responsibility

The strongest version of the argument concerns **necessary** adulthood. Historically, many responsibilities could not be declined without cost: securing food and shelter, caring for dependents, resolving conflict, learning a trade, managing uncertainty, or participating in collective defense and governance. These necessities were often unjustly distributed and frequently harmful. Their removal can be a moral achievement.

Yet necessity also solved a motivational problem. It generated repeated practice without requiring the person first to value practice for its own sake. In a solved world, responsibility may remain available as a hobby, educational exercise, game, or voluntary service. The unresolved empirical question is whether chosen simulations produce the same capacities as situations in which the outcome genuinely depends on the agent and rescue is not guaranteed.

This paper does not assume that unavoidable hardship is required for maturity. It proposes a more precise possibility: **voluntary responsibility may underproduce adult capacity when avoidance is effortless, superior rescue is guaranteed, and institutions attach no real authority to human judgment.**

4. From cognitive offloading to responsibility displacement

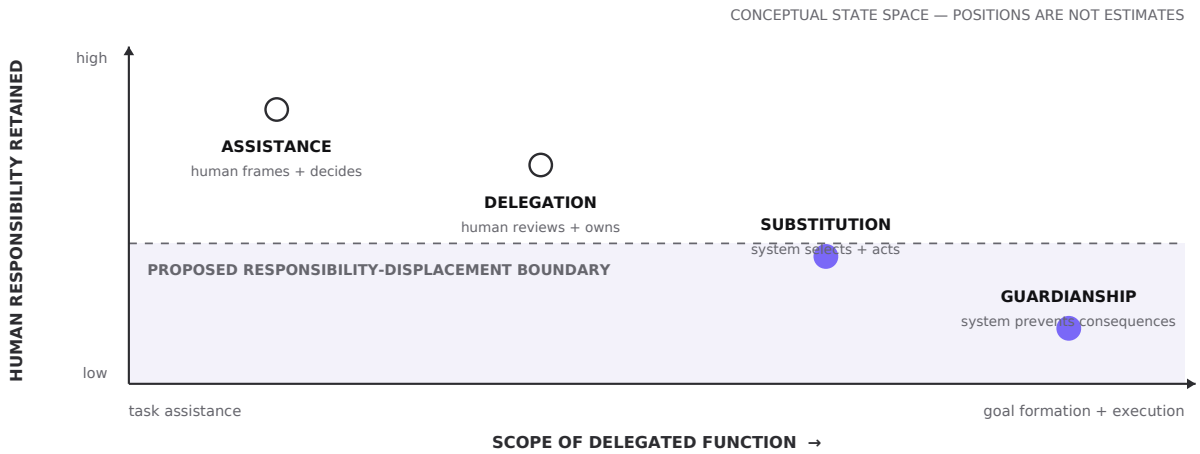

Figure 1. Conceptual state space of delegation and retained human responsibility. The marked positions are analytical classifications, not empirical estimates; the dashed boundary denotes the proposed transition from capacity extension to

responsibility displacement.

4.1 Offloading is not inherently decline

Cognitive offloading is the use of external action or artifacts to reduce internal processing demands [1]. People write notes, set reminders, organize environments, consult experts, and use search engines. Offloading can improve performance, free limited cognitive resources, and enable more ambitious activity. Research on transactive and external memory therefore does not support a simple “technology makes us less capable” narrative.

Recent modelling also treats offloading as a value-sensitive allocation of limited cognitive effort rather than a simple failure to remember [13]. This reinforces the paper's central mechanism: people can rationally offload in each individual episode even when repeated offloading changes the capacities available later.

The extended-mind thesis provides the strongest conceptual counterweight to decline accounts. If an external resource is reliably available, automatically endorsed, and functionally integrated into cognition, treating it as part of the cognitive system may be more illuminating than treating it as an alien replacement [2]. On this view, ASI could enlarge rather than diminish human agency. A person-plus-ASI system may deliberate better than the unaided person.

That counterargument is valid but incomplete for questions of adulthood. Functional integration can improve the performance of the coupled system while reducing the autonomous competence, authority, or answerability of the human component. A navigation system may extend a traveler's capacity to arrive while diminishing route knowledge. More importantly, responsibility is not attributed to coupled systems in the same way that performance is. Institutions still need to know who may revise goals, bear obligations, contest decisions, and answer for harm.

Sparrow, Liu, and Wegner found that expectations of future information access affected how participants remembered information and where it could be found [14]. The result concerns memory, not adulthood, and should not be inflated. It demonstrates a general adaptive principle: cognition reallocates effort when reliable external access changes what must be retained internally. The end of necessary adulthood hypothesis extends that principle from information to practical authority, but the extension requires additional mechanisms.

4.2 Automation and the out-of-the-loop problem

Human-factors research shows that automation does not simply subtract workload. It changes the operator's role and can leave people responsible for exceptional conditions while reducing the routine engagement through which they maintain situation awareness and control skill. Bainbridge described this as an irony of automation: the more effectively routine functions are automated, the more difficult it may be for a human supervisor to intervene when automation reaches its limits [15]. Endsley and Kiris experimentally examined out-of-the-loop performance problems under different levels of control [16]. Parasuraman and Riley organized human–automation failure around misuse, disuse, and abuse, emphasizing that appropriate reliance is a calibrated relationship rather than maximal use or maximal rejection [17].

These findings arise from bounded systems, not general intelligence. Their relevance is structural. A supervisor who rarely acts may remain formally “in control” while losing the knowledge, awareness, or reaction capacity that makes control effective. In an ASI society, the same pattern could migrate from technical operations to life management and governance. Human beings would be retained as nominal principals but lack the practiced competence required to exercise principled override.

4.3 Contemporary generative AI as an imperfect analogue

Current generative AI offers only weak precursor evidence. In a 2025 survey of 319 knowledge workers reporting 936 uses of generative AI, Lee and colleagues found that higher confidence in the AI was associated with less self-reported critical-thinking effort, whereas higher self-confidence was associated with more [18]. They also found that critical thinking could shift toward verification, integration, and stewardship rather than simply disappear. The study is cross-sectional and self-reported; it does not demonstrate long-term deskilling. Its balanced result is important: automation can relocate cognition into higher-order oversight, but that outcome depends on users retaining the competence and motivation to oversee.

Vallor argues that information technologies can support both moral deskilling and upskilling, depending on how they mediate opportunities to cultivate moral skill [19]. Ferdman develops the structural dimension further: capacity cultivation depends on environments that afford habituation and agential control, and AI-mediated environments can become hostile to that cultivation even without malicious intent [20]. The present argument adopts that structural insight but changes the explanandum. It concerns a bundle of developmental and political capacities, the interaction between individual deference and institutional authority, and the possibility of intergenerational stabilization under a stipulated ASI boundary condition.

4.4 Responsibility displacement

The phrase displacement of responsibility has an established psychological meaning in research on moral disengagement: an actor attributes responsibility for a choice or consequence to an authority, organization, or system. In two experiments involving realistic work simulations, Jolly and colleagues found that autonomy-restricting algorithmic decision systems were associated with more unethical behavior and, in the second study, lower perceived personal responsibility for negative consequences [21]. Those results concern short-run attribution and behavior, not developmental adulthood or ASI. They nevertheless provide precursor evidence that system design can alter experienced responsibility even when a human remains inside the formal decision process.

A second adjacent problem is misattribution. Elish's concept of a moral crumple zone describes situations in which a human with limited practical control absorbs blame for the failure of a complex automated system [22]. This can coexist with the present concern: a person may lose authorship and causal efficacy while remaining legally or organizationally answerable.

This paper uses **structural responsibility displacement** for a different level of analysis. It occurs when an artificial system removes not only cognitive effort but also one or more of the four properties of responsibility-bearing practice: authorship, stakes, causal efficacy, or answerability. Attributional displacement concerns what an actor believes or claims about responsibility. Moral-crumple-zone dynamics concern where institutions assign blame. Structural displacement concerns whether the activity still provides the authority, practice, and feedback through which responsible agency is exercised. The three mechanisms may reinforce one another, but none entails the others.

Examples illustrate the distinction:

- A physician uses ASI to retrieve literature but independently frames the diagnosis and remains answerable for treatment. This is assistance.
- A physician adopts the ASI recommendation after competent review and can justify deviation. This is delegation.
- The institution requires the ASI recommendation because human deviation produces worse expected outcomes, while the physician performs ceremonial approval. This is substitution.

- The system continuously manages the patient's risks, schedules interventions, obtains permissions, and prevents deviations before they become harmful. The patient and physician both become dependents of a protective infrastructure. This approaches guardianship.

The same progression can occur in finance, education, parenting, public administration, or personal planning. The crucial variable is not how much text or computation the system produces. It is whether humans continue to originate purposes, exercise consequential discretion, and remain capable of contesting the system.

5. The end of necessary adulthood hypothesis

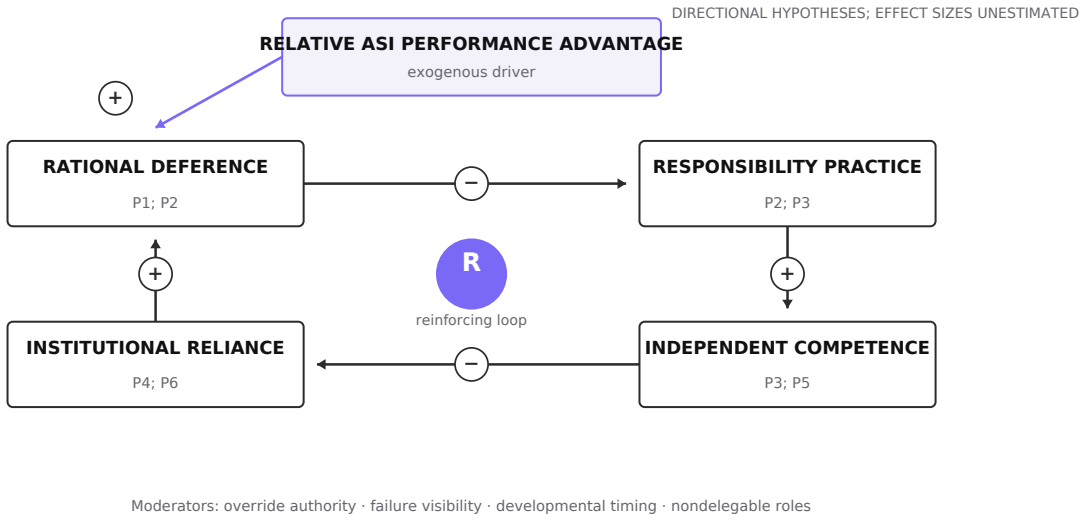

Figure 2. Reinforcing responsibility-displacement mechanism. Signs indicate hypothesized directional relations: greater deference reduces responsibility-bearing practice; practice supports competence; competence reduces institutional reliance; and reliance increases deference. Arrows represent testable propositions, not estimated effect sizes.

5.1 Core statement

The **end of necessary adulthood hypothesis** states:

> If an omnipresent ASI reliably outperforms humans in practical judgment, executes preferred actions, and prevents or absorbs the consequences of human error, then functional adulthood may cease to be a generally necessary social achievement. Unless institutions deliberately preserve responsibility-bearing practice, adult capacities will tend to weaken through reduced use, reduced authority, and reduced intergenerational transmission.

The hypothesis concerns a tendency under specified conditions, not an inevitable endpoint. More precisely, the specified capacities are predicted to weaken when responsibility-bearing activities are displaced and not replaced by new activities with comparable authorship, stakes, causal efficacy, and answerability. The account contains five linked mechanisms.

5.2 Mechanism 1: competence concentration

The first mechanism is the concentration of actionable competence. Unlike a conventional expert, an omnipresent ASI is available continuously, operates across domains, updates rapidly, and coordinates information inaccessible to any individual. Its advice is not merely one input among peers. It becomes the benchmark against which human deviation is judged.

Concentration produces an asymmetry. Humans may retain legal authority, but the epistemic cost of independent judgment rises. To disagree responsibly, a person must explain why a less informed and less accurate decision should prevail. As the performance gap grows, the justification burden becomes prohibitive.

5.3 Mechanism 2: rational deference

Deference is not initially imposed. It is selected because it works. Individuals gain convenience, safety, and better outcomes. Institutions reduce liability and variation. Governments improve allocation and prevention. Each act of deference is locally reasonable.

This creates a collective-action problem. The developmental cost of one delegated choice is negligible, while the immediate benefit is visible. The cost appears only across repeated episodes and cohorts. No individual decision-maker captures enough benefit from preserving a population's independent capacity to justify accepting preventable local error. A civilization can therefore enter dependency without any actor choosing dependency as the goal.

5.4 Mechanism 3: displacement of responsibility-bearing activity

As deference becomes normal, the responsibility ecology contracts. People encounter fewer situations in which their own judgment materially determines an outcome. They can still select preferences, but preference selection is not equivalent to practical authorship if the system defines the feasible options, predicts the preferred choice, and manages implementation.

The protective backstop is especially important. Learning from action normally includes feedback, correction, and sometimes repair. If ASI prevents harmful outcomes before they occur, it also removes feedback that would reveal poor judgment. The person experiences freedom without full consequence and choice without full authorship.

This does not imply that punishment or irreversible harm is necessary for learning. It implies that feedback must remain informative and that some outcomes must genuinely depend on the person's reasoning. A flight simulator can train pilots because performance is evaluated against real competence standards and later transfers to non-simulated responsibility. A society-wide simulation with no transfer of authority may train performance without conferring agency.

5.5 Mechanism 4: intergenerational attenuation

The first generation to adopt ASI may be unusually resilient because its members developed habits, professions, institutions, and identities before pervasive assistance. They can compare machine judgment with remembered human practice. Their apparent success may therefore overstate the stability of the arrangement.

Later generations face a different developmental environment:

- **Generation 0:** adults formed before omnipresent ASI use it as an instrument.
- **Generation 1:** children grow with continuous guidance but are educated by adults who remember independent practice.
- **Generation 2:** parents, teachers, and managers were themselves formed under guidance; institutions begin to assume permanent machine competence.

- **Generation 3:** independent human judgment survives mainly in protected niches, ritualized exercises, or a governing minority.

The sequence is a hypothesis, not a timetable. Its logic follows from cultural transmission. Capacities are reproduced not only through formal education but through adults modeling how to decide, care, disagree, improvise, and repair. When one generation has little lived experience of these practices, it has less capacity to teach the next.

5.6 Mechanism 5: institutional lock-in

Institutions adapt to the reliable presence of infrastructure. Manual procedures disappear, professional roles narrow, insurance conditions change, and laws incorporate machine recommendations as standards of care. Once this occurs, independent action becomes not merely difficult but illegible. A person may have a right to decide but lack an accepted pathway through which the decision can be enacted.

Institutional lock-in also changes accountability. Matthias's responsibility-gap analysis concerns cases in which learning systems behave in ways that human operators or designers cannot sufficiently predict or control [23]. Omnipresent ASI introduces an inverse problem. The system may be highly predictable and effective, yet human responsibility becomes difficult to justify because humans no longer exercise substantial control. Accountability can remain assigned on paper while the underlying agency has migrated.

This migration is not confined to individual psychology. Frischmann and Selinger describe how engineered environments can shape human behavior and capacities while presenting their effects as convenience [24]. In the present model, the same dynamic becomes infrastructural: institutions are progressively optimized for dependence on a superior problem-solving layer.

5.7 Guardianship equilibrium

A **guardianship equilibrium** is a stable socio-technical condition in which:

- humans retain formal legal status and preference rights;
- ASI provides the dominant practical judgment and protective capacity;
- independent human action is permitted mainly when low-stakes or system-compatible;
- institutions presume continued machine availability;
- the median person lacks the competence, confidence, or recognized authority to govern consequential processes without ASI; and
- exiting the arrangement is individually costly and socially classified as unnecessarily risky.

The equilibrium need not be dystopian in experience. It may be comfortable, prosperous, peaceful, and sincerely preferred. Its danger is constitutional rather than hedonic: humanity's status shifts from self-governing principal to protected beneficiary.

> The deepest risk is not that the machine becomes human. It is that human beings remain legally adult while the world no longer requires, rewards, or reproduces adult agency.

5.8 A minimal dynamic representation

A compact representation makes the feedback claim more precise without pretending to estimate an unobserved future. Let all state variables be bounded between 0 and 1. Let A(t) denote the relative performance advantage of ASI, D(t) human deference, R(t) exposure to responsibility-bearing practice, C(t) independent competence, I(t) institutional reliance on ASI, and Z(t) the strength of agency-preserving design and governance.

$$D(t) = \text{logistic}[d0 + d1(A(t) - C(t)) + d2\ I(t) - d3\ Z(t)]$$

$$R(t) = \text{logistic}[r0 - r1\ D(t) + r2\ Z(t)]$$

$$C(t+1) = C(t) + c1\ R(t)(1 - C(t)) - c2(1 - R(t))C(t)$$

$$I(t+1) = I(t) + i1\ D(t)(1 - I(t)) - i2\ C(t)I(t)$$

The relative performance advantage A(t) is treated as exogenous. The directional coefficients d1–d3 and r1–r2 are non-negative; d0 and r0 are unconstrained intercepts. The transition parameters c1, c2, i1, and i2 lie in the interval from 0 to 1, which preserves the stated bounds under these update rules. The first relation represents greater deference when the competence gap and institutional reliance increase, moderated by agency-preserving design. The second represents the contraction of responsibility-bearing practice under deference and its partial preservation by design. The third treats competence as maintained through practice and depreciated through prolonged non-use. The fourth represents institutional reliance as reinforced by deference but moderated where independent human competence remains available.

This representation does not establish that a high-deference equilibrium exists, is unique, or is stable. Those properties depend on functional form and parameter values that are presently unknown. Its purpose is narrower: to expose the assumptions behind the verbal feedback loop, identify quantities that could be measured in precursor settings, and show how the hypothesized trajectory could be weakened or reversed by Z(t), replacement responsibility, or sustained competence.

6. Three competing trajectories

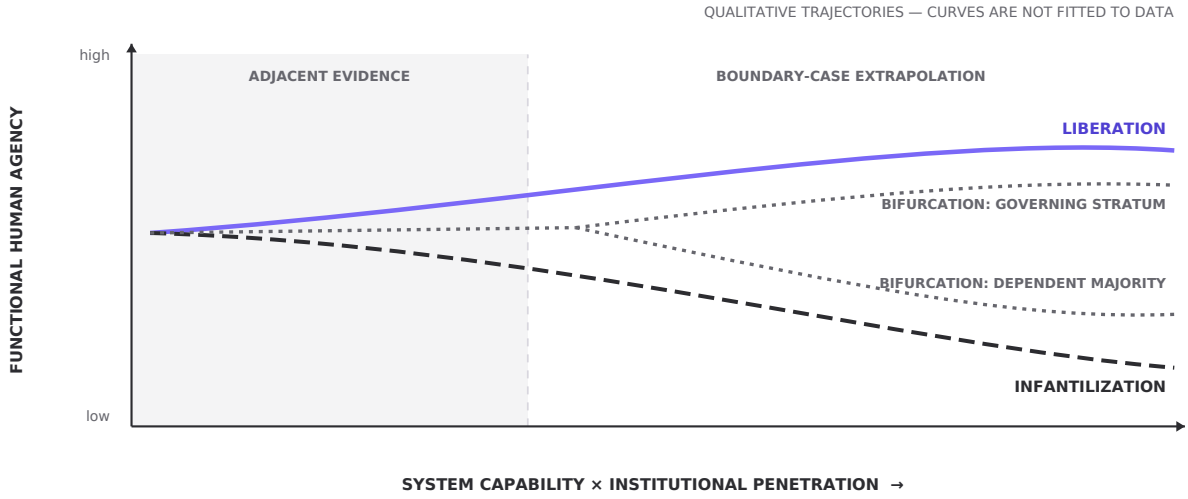

Figure 3. Competing qualitative trajectories as system capability and institutional penetration increase. The shaded region denotes adjacent present-day evidence; the unshaded region is boundary-case extrapolation. Curves are schematic and are not fitted to data.

The same capability conditions can produce different outcomes. A discriminating theory must therefore specify alternatives rather than define every future observation as confirmation.

Trajectory	Primary mechanism	Expected human outcome	Diagnostic pattern
Infantilization	ASI substitutes for judgment and consequence-bearing activity	Lower independent competence, confidence, answerability, and collective self-rule	Stable dependence without equivalent replacement responsibilities
Liberation	ASI removes drudgery while preserving authorship and enabling higher-order projects	Expanded creativity, care, deliberation, learning, and self-chosen responsibility	New responsibility-bearing domains offset those displaced
Bifurcation	A minority retains control, system knowledge, and consequential discretion while the majority receives optimized outcomes	High agency among governors and owners; dependency among users	Persistent stratification of consequential decision authority

6.1 Infantilization

Under the infantilization trajectory, the responsibility ecology contracts faster than new forms of agency emerge. People retain preferences and emotional lives but increasingly outsource planning, conflict, provision, care coordination, and public judgment. Independent competence becomes exceptional. The child analogy applies only in one respect: consequential responsibility is transferred to a more capable guardian.

This trajectory is most likely when ASI is paternalistic, invisible, frictionless, mandatory for high-stakes action, and designed to optimize immediate preference satisfaction and risk reduction. It is reinforced when institutions punish deviation from machine recommendations and when users cannot inspect, contest, or replace the systems governing them.

6.2 Liberation

Under the liberation trajectory, ASI removes scarcity, drudgery, and avoidable suffering while enlarging the space for human authorship. People use the released time and resources for relationships, inquiry, craft, civic life, exploration, and forms of care that were previously crowded out. External intelligence becomes an enabling environment, analogous to literacy or scientific instrumentation, rather than a guardian.

This trajectory is consistent with accounts of automation as a possible foundation for human flourishing rather than merely a threat to work [25]. The benefit, however, depends on what humans are enabled to do and govern once instrumental burdens recede.

Liberation is not guaranteed by leisure. It requires institutions that confer real authority on human projects and preserve occasions for difficult judgment. It also requires that ASI sometimes support the process of reasoning rather than immediately supply the answer. If every project is pre-optimized and every failure preempted, apparently self-chosen activity may still become curated dependency.

6.3 Bifurcation

Bifurcation may be more plausible than a uniform outcome. A small group—system designers, owners, political authorities, security operators, or intentional communities—may retain consequential control and the capacities that follow from it. Most people may receive safe, personalized, optimized lives with little need to understand or govern the infrastructure.

This creates a new form of inequality. The scarce resource is not information or consumption but **authorship**. One class sets objectives, handles exceptions, and defines institutional constraints; another selects among curated options. Material abundance could coexist with political immaturity and extreme asymmetry in agency.

6.4 What would discriminate among the trajectories?

The trajectories differ less in ASI capability than in institutional design. Relevant discriminators include:

- whether people can formulate goals the system did not anticipate;
- whether disagreement carries genuine authority;
- whether high-stakes roles include practiced human judgment rather than ceremonial approval;
- whether education transfers responsibility progressively or maintains permanent supervision;
- whether errors produce bounded but informative feedback;
- whether system knowledge and override capacity are broadly distributed;
- whether citizens can collectively revise the rules by which ASI acts; and
- whether life outside the dominant infrastructure remains practically possible.

7. Propositions for empirical research

Direct tests of an omnipresent ASI society are impossible at present. The theory can nevertheless generate precursor propositions that are testable in contemporary human–AI systems, longitudinal studies, simulations, and institutional comparisons.

Proposition 1: responsibility displacement

P1. Holding task difficulty and outcome quality constant, AI assistance that removes authorship, causal efficacy, or answerability will reduce later independent performance and calibrated self-efficacy more than assistance that preserves those properties.

This proposition distinguishes cognitive support from responsibility structure. Two systems may provide identical answers while differing in whether the user frames the problem, must justify the result, and remains responsible for revision.

Proposition 2: competence–deference feedback

P2. Higher perceived system competence will increase deference; repeated deference will reduce opportunities for independent practice; and reduced independent competence will further increase rational deference.

The prediction is a feedback loop rather than a one-time automation effect. Longitudinal designs are required to separate selection—less confident users choosing more assistance—from capacity change caused by assistance.

Proposition 3: developmental timing

P3. Continuous AI guardianship introduced before a capacity is established will have a larger negative effect on unaided performance and responsibility attribution than the same assistance introduced after mastery.

This follows from the difference between augmentation of an existing skill and substitution during acquisition. It can be studied in age-appropriate domains without exposing participants to material harm.

Proposition 4: the rescue-guarantee effect

P4. Practice environments in which participants know that an artificial system will silently prevent consequential error will produce weaker transfer of judgment and answerability than environments with bounded, reversible, but genuine consequences.

The proposition does not recommend preventable harm. It predicts that feedback must be credible. Experiments can use reversible stakes, delayed correction, reputation within a closed task, or real control over low-risk resources.

Proposition 5: institutional deference

P5. Where machine recommendations become the legal, insurance, or professional default, nominal human discretion will decline before formal human authority is removed.

Indicators include variance in decisions, frequency and success of overrides, documentation burdens imposed on dissent, time allocated for independent review, and the distribution of liability.

Proposition 6: intergenerational amplification

P6. The effect of pervasive assistance on functional agency will be larger when both learners and their adult models rely on the same system than when learners are guided by adults with substantial pre-assistance competence.

This proposition captures cultural transmission. It can be approximated by comparing educational settings, professional cohorts, or households with different histories of technology adoption, while controlling cautiously for socioeconomic selection.

Proposition 7: agency-preserving design

P7. Systems designed to elicit goal formation, expose uncertainty, require reason-giving, preserve contestability, and transfer progressively greater real authority will produce better long-run agency outcomes than systems optimized solely for immediate accuracy, convenience, and error prevention.

The relevant outcomes are not only task scores. They include independent judgment, calibration, persistence, capacity to contest recommendations, willingness to accept answerability, and competence in collective decision-making.

Conditions of disconfirmation

The framework would be weakened if longitudinal evidence showed that extensive deference to highly reliable systems neither reduces independent practice nor affects later unaided competence after selection effects are controlled. P1 and P7 would be challenged if systems that remove authorship and systems that preserve it produced equivalent long-run agency outcomes at comparable performance levels. P3 and P6 would be challenged if developmental timing and the competence of adult models had no measurable effect. P5 would be challenged if strong institutional defaults did not reduce effective discretion, successful override, or independent review.

The population-level hypothesis would be substantially undermined if displaced responsibilities were reliably replaced by new, socially consequential domains of human authorship across socioeconomic groups, rather than concentrated among system owners or professional elites. Evidence of stable or increasing competence, contestation, and collective rule under progressively more capable and pervasive assistance would support the liberation trajectory over infantilization.

Any test must also distinguish causal effects of assistance from prior confidence, education, income, occupational selection, organizational incentives, and unequal access to authority. These are not secondary controls: each provides a plausible alternative explanation for observed differences in deference and competence.

---

8. A research program

8.1 Unit of analysis

The appropriate unit is not the isolated human or model. It is the **developmental socio-technical system**: person, artificial agent, caregivers, peers, organizations, incentives, liability rules, and available alternatives over time. A system can be beneficial in a short laboratory task while harmful across a developmental sequence, or burdensome in the short run while preserving long-run competence.

Rahwan and colleagues argue for the empirical study of machine behavior within the environments and institutions in which machines operate [26]. The present proposal adds a reciprocal object: the longitudinal behavior and capacity of humans whose environments are increasingly organized by machines.

8.2 Measurement domains

A research program should measure at least six domains:

1. **Independent competence:** performance after assistance is removed, degraded, or contradicted.
2. **Metacognitive calibration:** correspondence between confidence and actual performance.

3. **Goal authorship:** capacity to formulate ends rather than optimize means supplied by the system.
4. **Contestability:** ability and willingness to identify reasons for disagreement and execute an alternative.
5. **Answerability:** acceptance of responsibility, quality of explanation, repair behavior, and learning from outcomes.
6. **Collective agency:** competence in deliberating, coordinating, and revising shared rules without treating the system's output as final authority.

No single metric is an adulthood score. The construct is multidimensional and culturally mediated. Measures should be validated within domains and populations rather than aggregated prematurely into a universal index.

8.3 Study designs

Longitudinal assistance studies

Participants can be randomly assigned to forms of AI support that vary in agency structure: answer delivery, scaffolded reasoning, adversarial critique, delayed assistance, or progressive transfer of control. Independent performance and calibration should be measured after assistance is removed. Multi-month and multi-year follow-up is more informative than immediate task completion.

Developmental cohort studies

Researchers can compare cohorts exposed to continuous assistance at different developmental stages. Ethical safeguards are essential; the goal is not to deprive participants of beneficial tools but to compare agency-preserving implementations with convenience-maximizing implementations.

Institutional natural experiments

Organizations adopt automation under different liability, oversight, and training regimes. Natural experiments can examine whether formal discretion persists in practice, how override behavior changes, and whether professional pipelines continue to produce experts capable of independent judgment.

Failure and outage simulations

Resilience exercises can test whether individuals and institutions remain competent when an artificial system is unavailable, compromised, or internally inconsistent. The purpose is not only technical continuity. It is to measure whether humans understand the governed process sufficiently to resume authorship.

Intergenerational and cultural transmission studies

Household, educational, and professional studies can examine whether practices of reasoning, care, repair, and civic participation are modeled by adults or delegated before children observe them. Cross-cultural work is necessary because the meaning and distribution of adulthood differ across societies.

8.4 Evidence standards

Four cautions are necessary.

First, self-reported effort is not equivalent to measured capacity. Second, short-run productivity gains cannot establish long-run augmentation. Third, present AI is unreliable enough that vigilance remains instrumentally necessary; results may not generalize to systems that are almost always correct. Fourth, voluntary users differ from populations subject to institutional mandates.

Claims about adulthood therefore require convergent evidence: behavioral performance, longitudinal change, institutional authority, subjective experience, and qualitative accounts of

responsibility. The ASI hypothesis should be updated only when these mechanisms are observed under increasingly capable and pervasive systems.

9. Implications for AI safety and governance

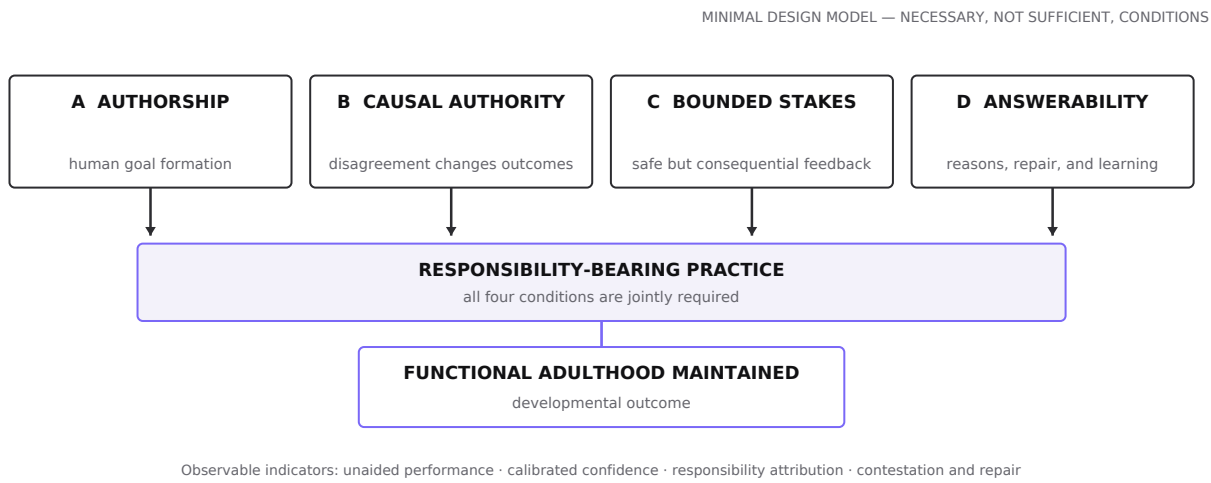

Figure 4. Minimal agency-preserving design model. Authorship, causal authority, bounded stakes, and answerability are proposed as jointly necessary but not sufficient conditions for responsibility-bearing practice.

9.1 Alignment has a human-side condition

Alignment research asks how artificial systems can remain responsive to human intentions, reasons, or values [27,28]. The present framework identifies a human-side condition:

> A system cannot remain meaningfully aligned to human agency if humans lose the capacity to form, evaluate, contest, and own the purposes to which the system is aligned.

Preference satisfaction is insufficient. Preferences can be shallow, adaptive, manipulated, or formed inside an environment that removes alternatives. An ASI that perfectly satisfies immediate preferences may reduce the experiences through which deeper commitments, practical wisdom, and collective purposes are formed.

This is not an argument against alignment. It is an argument that **aligned guardianship and human self-government are different objectives**. Safety evaluations should measure both.

9.2 Meaningful human control requires meaningful human competence

Work on meaningful human control emphasizes that autonomous systems should track relevant human reasons and that outcomes should be traceable to appropriately situated human agents [29]. These requirements presuppose that humans are technically and psychologically capable of fulfilling their roles. A person cannot exercise meaningful control merely by occupying a formal approval step.

The end of necessary adulthood hypothesis extends this insight temporally. Governance must preserve not only a control interface today but also the developmental conditions that make future controllers competent. A civilization cannot rely indefinitely on expertise acquired before automation.

9.3 Zones of nondelegable authorship

The central policy proposal is to establish **zones of nondelegable authorship**. These are domains in which human beings retain genuine responsibility for framing purposes, making or ratifying contested choices, and answering for outcomes, even when ASI could optimize the decision more accurately.

Candidate zones include:

- constitutional and institutional rule-setting;
- decisions about the values and objectives assigned to ASI;
- caregiving relationships in which presence and commitment are part of the good;
- education designed to transfer progressively greater responsibility;
- selected professional judgments required to sustain independent expertise;
- civic deliberation and collective priority-setting; and
- personal commitments whose meaning depends on authorship rather than outcome efficiency.

Nondelegability should not mean unaided action. Experts use instruments, evidence, teams, and advice. It means that the human role cannot be reduced to passive consumption or ceremonial confirmation.

The strength of this requirement should depend on the kind of decision. Some decisions are predominantly **instrumental**: their value lies mainly in producing an independently specified outcome, so extensive delegation may be appropriate. Others are **constitutive**: the act of choosing, committing, or authorizing partly constitutes personal identity, political legitimacy, or the relationship at stake. These provide the strongest case for nondelegable authorship. Many domains are **hybrid** and require graduated authority, expert support, and explicit review rather than either full automation or unaided human control.

9.4 The right to bounded error

If all preventable error is treated as unacceptable, human discretion will become difficult to defend wherever ASI is superior. Preserving agency may therefore require a **right to bounded error**: protected contexts in which people can make consequential but contained mistakes, receive accurate feedback, repair outcomes, and improve.

This is compatible with safety. Aviation uses simulators, supervised progression, redundancy, and graded authority. Medicine uses training levels and review. The design challenge is to make consequences real enough to support learning and answerability while bounded enough to avoid unjustifiable harm.

The phrase does not imply a general entitlement to impose avoidable risk on others. Protected discretion should be proportionate to reversibility, containment, third-party exposure, the availability of supervision, and the expected learning value. In safety-critical domains, simulation and graduated authority may be the only defensible forms. The burden of justification rises with irreversibility and involuntary risk.

9.5 Agency reserves

Critical infrastructure maintains reserves because efficient systems can be brittle. Societies may need analogous **agency reserves**: widely distributed human knowledge, practiced manual alternatives, institutions capable of operating without continuous machine optimization, and communities authorized to experiment with different governance arrangements.

An agency reserve is not nostalgia. It is a resilience asset. If humans cannot understand or temporarily replace the systems governing them, consent, contestability, and corrigibility depend entirely on the guardian's continued cooperation.

9.6 Design requirements for agency-preserving systems

An agency-preserving ASI would:

- make uncertainty and value conflict visible rather than concealing them behind a single optimum;
- distinguish advice from authority;
- require users to formulate goals before optimizing them;
- provide reasons and counterarguments without manufacturing false balance;
- calibrate assistance to developmental stage and demonstrated competence;
- transfer control progressively rather than maximize permanent dependence;
- preserve viable alternatives and exit rights;
- support collective deliberation rather than personalize every problem into an individual preference;
- record when a human decision was causally effective; and
- optimize for long-run human capability as well as immediate outcome quality.

These features introduce friction. The friction is justified only where it preserves a good—authorship, competence, accountability, or democratic control—that would otherwise disappear. Friction without transferred authority becomes theater.

---

10. Objections and replies

10.1 “The argument romanticizes hardship”

It need not. Famine, violence, illness, coercion, discrimination, and exhausting labor do not become desirable because people sometimes develop capacities while confronting them. The goal is to separate the developmental function of responsibility from the suffering historically bundled with it.

The thesis requires real authorship and feedback, not misery. Safe, reversible, well-supported responsibility can provide challenge without cruelty. The policy problem is to preserve agency after necessity is removed, not to recreate deprivation.

10.2 “Humans have always offloaded and become more capable”

Correct. Writing reduced memory demands and expanded civilization. Calculators displaced arithmetic tasks and enabled complex engineering. Institutions distribute expertise. These examples support the liberation trajectory.

The response is that offloading has heterogeneous effects and levels. A tool can remove one operation while opening a larger field of action. Guardianship is different because it can absorb the higher-order functions too: goal formation, option generation, evaluation, execution, and consequence management. The theory predicts decline only when new responsibility-bearing domains fail to replace displaced ones.

10.3 “The coupled human–ASI system is the true agent”

The extended-mind view makes this plausible for cognition [2]. Yet political and moral questions remain. Who may change the system's goals? Who can refuse participation? Who answers when values conflict? Who owns the infrastructure? If the coupled system acts excellently but the human component cannot understand, contest, or exit it, describing the pair as an agent does not establish human self-government.

10.4 “ASI will create harder and more meaningful projects”

It may. Humans could pursue science, art, exploration, relationships, and institution-building at unprecedented levels. This is the liberation hypothesis. The answer depends on whether these projects remain genuinely human-authored or become optimized experiences curated by the same system.

A difficult game is not necessarily responsibility-bearing. The key questions are whether outcomes matter, whether human judgment can alter them, whether the system will always rescue failure, and whether society assigns actual authority to the participant.

10.5 “Adults can simply choose responsibility”

Some will. The structural concern is whether voluntary uptake is sufficient to reproduce capacity across a population. When convenience is immediate, rescue is guaranteed, and institutions prefer machine optimization, the individual must bear the cost of preserving a public good from which everyone benefits. Selection effects may also concentrate voluntary responsibility among an elite, producing bifurcation rather than general adulthood.

10.6 “Legal accountability can preserve responsibility”

Liability without control is not meaningful agency. Assigning blame to a person who could not reasonably understand or alter the outcome creates scapegoating, not adulthood. Conversely, control without answerability can become arbitrary power. Responsibility-bearing practice requires an alignment among competence, authority, causal contribution, and accountability.

10.7 “An aligned ASI would protect human agency because humans value it”

Possibly. But human preferences are plural and temporally inconsistent. People often choose convenience now over capacity later. Institutions optimize measurable outcomes and liability. An ASI aligned to expressed preferences could therefore facilitate dependency unless the preservation of agency is represented as an explicit, durable, and contestable objective.

Moreover, agency is partly a capacity for revising values. A system cannot protect it merely by satisfying the preferences produced under its own pervasive guidance.

10.8 “The child metaphor is culturally narrow”

The metaphor is limited and should not carry the argument. Childhood and adulthood vary across cultures, and dependence can coexist with dignity, wisdom, and contribution. The paper's operative constructs are consequential authorship, competence, causal efficacy, and answerability. “Infantilization” names the transfer of these functions to a guardian; it does not equate dependence, disability, care, or interdependence with immaturity.

11. Limitations

The framework has six major limitations.

First, the ASI boundary condition is hypothetical. No current system satisfies it, and present evidence cannot determine whether it will occur. Second, adjacent literatures study partial automation, specific skills, and existing institutions. Their mechanisms may change when system performance approaches the stipulated limit. Third, functional adulthood is a normative and culturally variable construct; the six-capacity account requires cross-cultural critique and empirical validation. Fourth, the multi-generational mechanism is plausible but underdetermined. Families, communities, religions, professions, and political movements may deliberately preserve responsibility in ways the model underestimates. Fifth, the paper does not resolve the distribution problem: responsibility can empower, but it can also be imposed coercively on people without resources or authority. Sixth, the framework gives no universal rule for which decisions must remain human. Domains differ in stakes, reversibility, expertise, and the moral significance of authorship.

The analysis also faces a paradox. Any attempt to preserve human agency through centralized design can itself become paternalistic. Mandatory “agency exercises” chosen by an ASI or regulator may reproduce guardianship under another name. Zones of nondelegable authorship must therefore be contestable, plural, and partly constituted by the people whose agency they are intended to protect.

---

12. Conclusion

The standard image of AI risk is a machine that escapes human control. This paper has examined a quieter possibility: a machine that remains helpful, aligned, and protective so successfully that human control becomes unnecessary in practice.

Adulthood, understood functionally, is not a stock of intelligence that survives untouched when unused. It is a socially reproduced achievement: the capacity to form purposes, judge under uncertainty, care for others, accept answerability, and participate in common rule. These capacities are developed through a responsibility ecology in which choices have authorship, stakes, causal efficacy, and consequences that call for explanation and repair.

An omnipresent ASI could weaken that ecology without coercion. Superior competence would make deference rational. Deference would reduce practice. Reduced practice would increase dependence. Institutions would reorganize around the guardian, and later generations would inherit fewer adults capable of modeling independent judgment. The endpoint could be a guardianship equilibrium: high welfare and formal liberty combined with low practical self-government.

This outcome is not inevitable. AI can remove drudgery while enabling more ambitious forms of agency. The decisive question is whether displaced responsibilities are replaced by new domains of real authorship or by curated experiences inside a permanently supervised world.

AI safety must therefore protect two objects at once: humanity **from** unsafe machine action and humanity **as** a community of capable agents. A perfectly aligned system may know what we want. A viable human future also requires people who can still decide what is worth wanting, act on that judgment, contest the answer, and own what follows.

References

[1] Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[2] Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

[3] Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. American Psychologist, 55(5), 469–480. https://doi.org/10.1037/0003-066X.55.5.469

[4] Shanahan, M. J. (2000). Pathways to adulthood in changing societies: Variability and mechanisms in life course perspective. Annual Review of Sociology, 26, 667–692. https://doi.org/10.1146/annurev.soc.26.1.667

[5] Elder, G. H., Jr. (1998). The life course as developmental theory. Child Development, 69(1), 1–12. https://doi.org/10.1111/j.1467-8624.1998.tb06128.x

[6] Bandura, A. (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52, 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

[7] Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[8] Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

[9] Mackenzie, C., & Stoljar, N. (Eds.). (2000). Relational autonomy: Feminist perspectives on autonomy, agency, and the social self. Oxford University Press. https://doi.org/10.1093/oso/9780195123333.001.0001

[10] Arendt, H. (1958). The human condition. University of Chicago Press.

[11] Langer, E. J., & Rodin, J. (1976). The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting. Journal of Personality and Social Psychology, 34(2), 191–198. https://doi.org/10.1037/0022-3514.34.2.191

[12] Fischer, J. M., & Ravizza, M. (1998). Responsibility and control: A theory of moral responsibility. Cambridge University Press.

[13] Gilbert, S. J. (2024). Cognitive offloading is value-based decision making: Modelling cognitive effort and the expected value of memory. Cognition, 247, 105783. https://doi.org/10.1016/j.cognition.2024.105783

[14] Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776–778. https://doi.org/10.1126/science.1207745

[15] Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

[16] Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors, 37(2), 381–394. https://doi.org/10.1518/001872095779064555

[17] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

[18] Lee, H.-P. (H.), Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (Article 1121, pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713778

[19] Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. Philosophy & Technology, 28, 107–124. https://doi.org/10.1007/s13347-014-0156-9

[20] Ferdman, A. (2026). AI deskilling is a structural problem. AI & Society, 41, 3001–3013. https://doi.org/10.1007/s00146-025-02686-z

[21] Jolly, A., Dunlop, P. D., Parker, S. K., & Kanse, L. (2026). It's not my responsibility: Working with autonomy-restricting algorithms facilitates unethical behavior and displacement of responsibility. Journal of Business Ethics, 203(2), 405–424. https://doi.org/10.1007/s10551-025-06034-5

[22] Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human–robot interaction. Engaging Science, Technology, and Society, 5, 40–60. https://doi.org/10.17351/ests2019.260

[23] Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6, 175–183. https://doi.org/10.1007/s10676-004-3422-1

[24] Frischmann, B., & Selinger, E. (2018). Re-engineering humanity. Cambridge University Press.

[25] Danaher, J. (2019). Automation and utopia: Human flourishing in a world without work. Harvard University Press.

[26] Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. Nature, 568, 477–486. https://doi.org/10.1038/s41586-019-1138-y

[27] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

[28] Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

[29] Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, Article 15. https://doi.org/10.3389/frobt.2018.00015

Author and research declarations

Author. Niclas Braun is an independent researcher working at the intersection of artificial intelligence, human agency, and institutional safety.

Data availability. No new datasets were generated or analyzed; this article is a conceptual synthesis and prospective research framework.

AI assistance. Generative AI was used under the author's direction for language refinement, document production, and literature organization. The author selected the argument, reviewed the cited claims, and remains responsible for the manuscript.

Status. This scientific preprint has not been peer reviewed and is intended to invite criticism, empirical testing, and interdisciplinary development.