Constitutive Pluralism:

Relational Value, Standing, and the Pace of AI-Mediated Change

Koichi Takahashi
Keio University, Japan
AI Alignment Network (ALIGN), Tokyo, Japan
shafi@keio.jp

Preprint, version 1, 9 November 2026

Abstract

AI systems can improve decisions while reducing the ability of those affected to question their purposes and conditions. Artificial general intelligence may erode supports of people's standing in decisions, such as the indispensability of labor. This paper develops constitutive pluralism, a principle that protects the conditions under which those who can suffer meaningful loss from decisions can question and choose again. People develop the ability to question in relationships and encounters with others, and institutions cannot do this for them. On grounds of equal consideration and the justification owed to those affected, the principle rules out fixing the content and ranking of value for a whole society by a single subject and a single measure, beyond effective contestation. Two foundations answer the questions this leaves open. Relational value locates some value in historically constituted processes of interaction, and vulnerability grounds who has standing in decisions that set value. Three pillars defend the foundations against the unification of the agent, measure, and reach of a single optimization process. Non-domination keeps power over relationships contestable, non-aggregation forbids binding comprehensive evaluations of persons, and slack preserves room that optimization does not occupy. Autonomy of pace protects the ability of those affected to take part in deciding when arrangements change or continue. Institutions that rely on current satisfaction and articulated claims can overlook eroding opportunities for relational value. The paper connects the principle to error correction in societies and adaptive capacity in civilization and revises an argument from a Japanese book.

Keywords: constitutive pluralism, relational value, standing, non-domination, autonomy of pace, AI governance

Contents

1 Introduction **4**

Part I. The principle **8**

2 The normative premise and what must not be decided **8**

3 Two foundations **10**
3.1 Where value arises: relational value 10
3.2 Who takes part in setting value: vulnerability and standing 15
3.3 The two foundations are not reducible to each other 19

4 Residual disagreement and the threshold of standing **20**

5 From the threat to three pillars **23**
5.1 Three loci of a single optimization process 23
5.2 Non-domination 24
5.3 Non-aggregation 25
5.4 Slack 28
5.5 Complementarity, functions, and means 29

6 Autonomy of pace **29**

7 Adaptive capacity at the scale of civilization **30**

8 Premises and what remains if they are rejected **32**

Part II. Institutions **32**

9 Comparing institutional arrangements **32**
9.1 Beneficial provision and control over purposes 33
9.2 Many providers and one comprehensive ranking 34
9.3 Public benefit from limited information linkage 35
9.4 Cumulative updates without a single controller 35
9.5 Continuation, change, and exit 36
9.6 Emergency action and the cost of waiting 37
9.7 Catastrophic risk and coordinated action 38
9.8 An illustration from the production of knowledge 38

10 Conditions of realization **38**

Part III. Testability **40**

11 Testing and revision **40**
11.1 Empirical claims and evaluation 40
11.2 Conditions under which plurality supports adaptation 40
11.3 Failure modes and evidence for reconsideration 41
11.4 Counterexamples and the scope of revision 42

12 Relation to existing theories and to the book **43**
12.1 Relation to existing theories 43
12.2 Relation to the book 45

13 Scope and conclusion **45**

Use of AI in the preparation of this paper **46**

References **46**

1 Introduction

Consider a public AI service that improves access to healthcare, education, and income support. Its recommendations are accurate. An independent auditor can correct mistaken eligibility decisions. However, the people who depend on the service cannot challenge which forms of education it makes available, which relationships it encourages, or which purposes guide its recommendations. They can obtain a better answer to an accepted question, but they cannot effectively question how the available questions and choices were determined. Better outcomes provide reasons to use the service. Even when the service is benevolent and technically competent, they do not by themselves settle who should determine its purposes or which objections it must recognize.

The concern extends beyond a single decision maker. A household in which one member cares for another may face simultaneous changes in its workplace, school, insurance, and public services. Taken separately, each provider's notice period may be reasonable. Together, the changes may leave that caregiver unable to understand them or obtain assistance before new conditions take effect. There need be no common operator, comprehensive plan, or deliberate attempt to exclude that person. The loss arises through the interaction of individually defensible arrangements.

These concerns become pressing with artificial general intelligence (AGI), an anticipated form of AI with broad competence across intellectual tasks. AGI does not abolish limits but moves them, so that limits derived from human cognition recede while physical, computational, relational, and institutional limits come to the fore (Takahashi 2026a). Three changes may follow.[1] First, when AI can replace human labor across much of the economy, it erodes the economic and military indispensability of workers, which has been a material support of modern rights and freedoms. Second, compute, data, and talent concentrate for reasons rooted in the technology itself, so that who owns and governs AGI becomes a central political question (Takahashi 2026a, Chapter 10). Third, as material goods and intellectual outputs become abundant, meaning, purpose, belonging, trust, and relationships become relatively scarce, and how they come about becomes a central question. Where these changes occur, each weakens a support on which people's claim to a say in the decisions that affect them has rested, and better outcomes alone do not restore it.

I propose constitutive pluralism as a principle for these conditions. It protects the conditions under which those who can suffer meaningful loss from decisions can question and choose again. Section 2 states this normative premise and defends it on grounds of equal consideration and the justification owed to those affected. The premise has a protective face, which concerns the conditions of questioning and choosing, and a standing face, which concerns taking part in decisions whose consequences one receives. Here, standing means a recognized claim to protection, reasons, contestation, and appropriate representation.[2] The premise rules out one kind of decision. Such a decision fixes the content and ranking of value for a whole society by a single subject and a single measure and binds outsiders and future participants to it without effective contestation.

The principle adopts a view of how the ability to question develops. People develop it in relationships in which reasons are asked for and given, in encounters with others whose responses are not subordinated to their expectations, and through access to contexts other than the one being questioned. Institutions and technologies shape these relationships and opportunities, but they cannot accomplish

[1] This paper treats these changes as conditional scenarios, because each is mediated by ownership, institutions, and the distribution of income.

[2] Standing in this sense is broader than standing to sue in a court.

this development on people's behalf. Institutions can secure the conditions under which it continues, and the principle addresses those conditions.

The name of the principle states its central claim. Pluralism refers to the plurality of values, reasons, relationships, and contexts that people can form and sustain, not to a count of providers. Constitutive means that this plurality is part of what the protected conditions consist in, not only a desirable feature of a society. Because the ability to question develops through access to other relationships and contexts, a society without such plurality lacks part of what the normative premise protects, whatever its other merits. The constitutive sense resembles the difference between valuing biodiversity and recognizing it as a condition for ecosystems to function (Takahashi 2026a, Chapter 11). Section 12.1 distinguishes the principle from other pluralisms in political theory and from appeals to plurality in AI and technology.

The prohibition leaves open where value arises and who has standing in decisions that set value. The two foundations of relational value and vulnerability answer these questions respectively (Section 3). A single optimization process has three loci: who carries it out, by what measure it evaluates, and how far it reaches. The three pillars of non-domination, non-aggregation, and slack defend these foundations against the unification of these loci (Section 5). Non-domination keeps authority over relationships and decisions under the effective control of those affected. Non-aggregation forbids the binding use of a comprehensive cross-context evaluation of a person, while it permits limited-purpose linkage and aggregation under conditions. Slack preserves time, resources, and discretion that optimization does not occupy. Autonomy of pace, the temporal face of the normative premise, protects people's ability to take part in deciding when arrangements change or continue (Section 6). Figure 1 summarizes this structure. The public service above lacks non-domination, because those who depend on it cannot contest its purposes. The household case lacks slack and autonomy of pace, because the combined changes leave the caregiver neither time nor opportunity to respond.

The principle has related but distinct purposes at three scales. At the scale of persons, it protects the conditions under which new meaning and value can arise between people whose contributions are not wholly subordinated to one another's expectations. It also protects the conditions under which a person's own questions remain open to change. At the scale of society, it aims to prevent any single objective function (an evaluative formula that specifies what is to be increased and by how much), measure of evaluation, or governing subject from acquiring final and binding authority over social life while foreclosing effective contestation of standing, purposes, and boundaries. A common measure is not objectionable merely because it is common. At the scale of human civilization, it aims to preserve the capacity to adapt to shocks whose kind, timing, or course cannot be foreseen, such as an asteroid impact, contact with a non-human intelligence, or a transformative technology such as AGI itself (Section 7). It does so by keeping a single set of assumptions from dominating the cognitive base through which humanity understands the world. These purposes apply one principle: no single framework should close the space in which others can respond. The normative ground lies at the scale of persons, and the larger scales supply further reasons, each needing its own support and each usable by a reader who rejects another (Table 1).

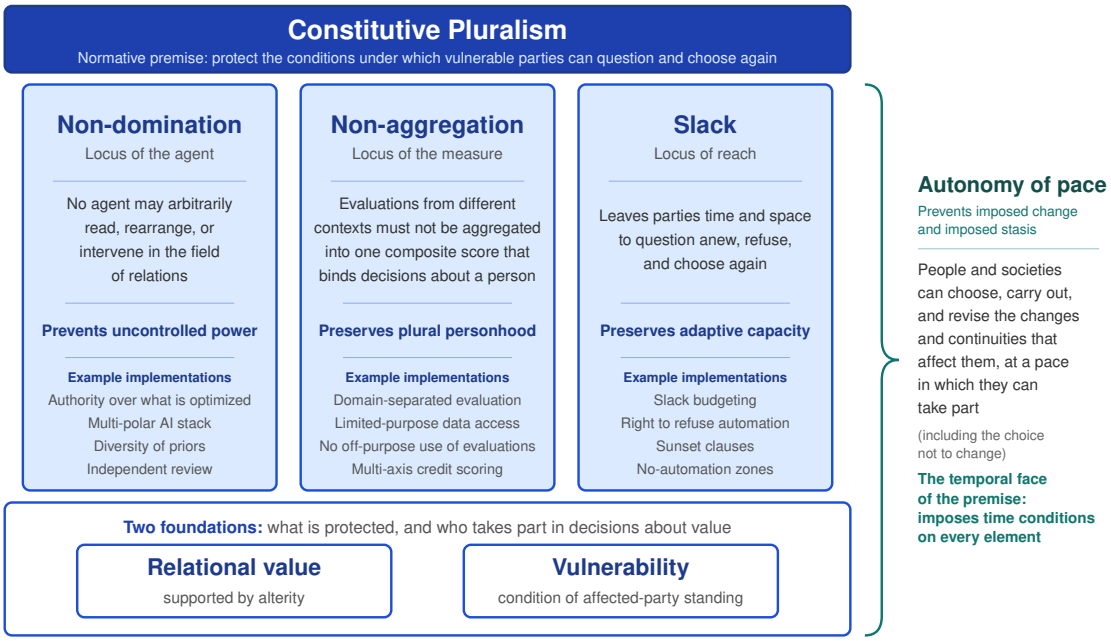


Figure 1: The structure of constitutive pluralism. The beam states the normative premise (Section 2). The two foundations answer where value arises (the question of genesis) with relational value and who takes part in setting value (the question of legitimacy) with vulnerability (Section 3). The three pillars guard against the unification of the three loci of a single optimization process (Section 5). Autonomy of pace, shown by the brace, is the temporal face of the normative premise rather than a fourth pillar, and it constrains the foundations, the pillars, and the institutions (Section 6). The example implementations are candidate means rather than requirements (Section 5.5). Alterity, the resistance, unpredictability, and uncontrollability of the other party, supports the relational value through which people discover or revise what they value, and this paper reconstructs it as independence, in which the other's contribution is not wholly subordinated to fulfilling one party's expectations (Section 3.1). The figure revises the one in the supplementary essay in which I developed the book's principle (Takahashi 2026b).

Table 1: The purposes of the principle at three scales and the status of their reasons.

Scale	What the principle protects	Status of the reason	Sections
Persons	Relationships in which value arises between people and in which questions can be shaken, and the standing of those who bear the consequences of decisions	The normative ground, which rests on equal consideration and the justification owed to those affected	2–4
Society	That no single objective function, measure, or governing subject acquires final and binding authority while foreclosing contestation	A further reason. The prohibition rests on the normative premise, and plurality can support the correction of error under suitable institutional conditions	2, 5, 6, 9
Civilization	The reservoir of frameworks and the capacity to generate new ones, on which adaptation to unforeseeable shocks can depend	A further reason, which is empirical and conditional on how errors are correlated, coordination, institutions that pass on what is learned, and containment	7, 11.2

The argument applies when AI materially shapes environments of choice and preference formation, determines evaluations and allocations across several domains, or drives cumulative changes that exceed the capacity of affected people and institutions to respond. These conditions can hold for narrower systems as well as for AGI. The argument therefore depends on neither a date of arrival for AGI nor a capability threshold.

I introduced the principle in a Japanese book on AGI (Takahashi 2026a) and developed it in a supplementary essay (Takahashi 2026b). The present paper reconstructs its derivation and justification as a self-contained argument. Section 12.2 states where the paper clarifies or revises the book. Many of the protections that follow have precedents in theories of liberty, non-domination, privacy, and governance (Section 12). Where those theories support the same protections for the same reasons, the paper's contribution is inheritance and synthesis. Its distinctive contribution is the arrangement itself, together with the argument for it. The arrangement grounds institutional protections in relational value and vulnerability, derives three pillars from the loci of a single optimization process, adds a temporal principle, and connects the three scales.

The paper has three parts. Part I (Sections 2 to 8) states the principle and the premises on which it rests, Part II (Sections 9 and 10) turns to institutions, and Part III (Section 11) sets out how its claims can be tested and what evidence would require revising them. After the three parts, Section 12 relates the principle to existing theories and to the book, and Section 13 concludes.

Part I. The principle

2 The normative premise and what must not be decided

The principle rests on a normative premise: the conditions under which vulnerable parties can question and choose again should be protected. Vulnerable parties are those who can suffer meaningful loss (Section 3.2). The premise has two faces. Its protective face requires that the conditions for raising questions, refusing, and choosing again not be impaired. Its standing face requires that those who bear the consequences of a judgment have standing to take part in it, directly or through representation. This face directly guarantees the core of standing, which consists of protection, reasons, and a route to independent review (Section 3.2). The form and extent of participation in collective decisions, and any specific power to decide, need further reasons. The two faces are connected. For those who bear the consequences of other people's judgments, being able to question and choose again includes being able to contest those judgments and to be involved in them (Takahashi 2026b, Section 1).

Two commitments give reasons for adopting the premise (Takahashi 2026b, Sections 1 and 11). The first is equal consideration, under which each affected person counts as an equal claimant whose interests must enter the justification, although not every interest has the same weight. The second is the justification owed to those affected. Decisions that bind others must be justified to them through an actual route by which reasons can be demanded and examined, not only impersonally. Both commitments are normative starting points that do not follow from any description of intelligence.

These commitments limit what competence can settle. A comparative advantage in competence, including that of a centralized arrangement, gives a reason to delegate some judgments, but it does not confer jurisdiction over the ends of the delegation. A system may select treatments better than a physician under an agreed objective without acquiring final authority over which burdens a patient should accept. Accurate predictions of present preferences likewise do not show that the conditions under which preferences develop should be beyond challenge. The argument thus calls for an account of authority with effective routes to dispute the scope, purposes, and continuation of a delegation.[3]

A further reason concerns who can demand justification. If an authority alone decides which objections count and whether they have been answered, consideration of a person's reasons depends on that authority's discretion. Reciprocal justification, in which those bound by a decision or their representatives can demand and examine its reasons, avoids this dependence. The position is close to the political philosopher Rainer Forst's account of a basic right to justification, in that both treat each person as a subject who can demand and examine justifications. Forst argues that the impulse against injustice is primarily a refusal to be dominated rather than a desire for more goods, and that the core issue of justice is "who determines who receives what" (Forst 2014, p. 8). The premise shares his distinction between receiving goods and being a subject of justification. These reasons do not bind those who reject both commitments, and facts about a decision maker's benevolence or competence cannot settle that normative disagreement (Section 4).

What must not be decided has one logical form, which combines two elements. In the first, a

[3]The text's distinction between competence and jurisdiction is compatible with the legal philosopher Joseph Raz's service conception of authority. That conception asks whether following an authority improves a person's conformity to the reasons that apply to that person. It also asks whether, in the matter concerned, such conformity should take priority over deciding without the authority's assistance (Raz 2006, pp. 1014–1016).

single subject fixes the content and ranking of value for a whole society by a single measure. In the second, that decision binds outsiders and future participants without effective contestation of standing, purposes, and boundaries. The combination is the most extreme form of the closure that the principle aims to prevent at the scale of society (Section 1). Individuals and communities may set values in their own contexts, which is what plurality consists in. A voluntary community, such as a religious one, may also share a comprehensive ranking internally. The ranking comes into question when members cannot object or cannot leave, or when it binds outsiders or future participants. The prohibition rests on the loss of the conditions that the premise protects. A common decision with a limited purpose therefore falls outside the prohibition, provided that those affected can effectively contest standing in it, its purpose, and its boundaries. In this paper, revisable means open to such contestation, not merely equipped with a formal route of revision. The binding use of a comprehensive evaluation of a person across contexts is a different form of closure, against which Section 5 gives separate reasons.

The prohibited decision concerns the first order of value, its content, such as what a good life is and what deserves priority. Across society as a whole, constitutive pluralism requires only the second order, the constitutive conditions under which that content can be contested and chosen again. Protecting these conditions is a substantive judgment that gives priority to equal standing, effective choice, and relational opportunity, and is thinner than a conception of the good but not empty. The constitutive conditions are themselves open to revision through the routes they guarantee. A revision may not, however, unilaterally erase the standing of future others to question and choose again (Takahashi 2026b, Section 2).

Plurality is also supported by the limits of what any agent can know and predict about a society, the limits of knowability (Takahashi 2026a, Chapters 7, 8, and 11). Four constraints remain however far the capability of AI rises. The first is the intrinsic time of the system under study, the time that the system itself takes to proceed, such as a cell cycle or the generational time of institutional change. No technology can shorten the intrinsic time of a given system. Technology can change the system, but the changed system then has an intrinsic time of its own. The second is information that cannot in principle be obtained, such as a history that has been lost or what quantum indeterminacy leaves open. The third is the indeterminability of value, which means that value itself cannot be settled by experiment.[4] The fourth is the limits of computation, which mean that some problems, such as deciding in general whether a program halts, cannot in principle be solved by computation. These limits hold for humans and for any AGI, and every real agent also acts with finite information, computation, and time. The limits rule out an unconditional guarantee that the long-term consequences of social decisions can be completely predicted and optimized under a single objective function. They therefore block the inference from superior capability to final and comprehensive authority over what a society pursues. The limits do not rule out comparative advantages of centralized arrangements under specified conditions, which are a matter for empirical comparison (Sections 9 and 11). Nor do they by themselves yield the prohibition, since an authority can still fix in fact what a society pursues. The force of the prohibition therefore comes from the normative premise.

This paper draws on the economist and social philosopher Friedrich Hayek's argument that no

[4]The book cites Kenneth Arrow's impossibility theorem (Arrow 1951) at this point (Takahashi 2026a, Chapter 8), and online Appendix 2 narrows the citation (Takahashi 2026c). The theorem concerns the aggregation of several persons' preferences, and it is conditional. The claim in the text rests instead on the point that converting distinct dimensions of value into one score requires choosing a rate of conversion, which is already a value judgment (Takahashi 2026b, Section 2).

central planner can gather dispersed knowledge (Hayek 1945, 1960). A principle that prescribes institutions may therefore seem to repeat the deliberate design of social order as a whole that Hayek warned against. The separation of evaluation across domains, veto rights, approval of exceptions, and sunset clauses are, however, boundary conditions that preserve what cannot be planned, not a plan for society as a whole (Takahashi 2026c). A sunset clause makes a rule lapse unless it is renewed. These boundary conditions extend Hayek's defense of spontaneous order, the order that arises from people's interactions without a designer of the whole. Once AI can aggregate dispersed information, explicit design must protect this order. Whether a concrete design of these conditions turns into the comprehensive control of social order that Hayek warned against depends on how it is carried out (Takahashi 2026c).

3 Two foundations

The prohibition stated in Section 2 does not settle what value is. Two questions nevertheless require answers. The first is the question of genesis, which asks what institutions should protect as the place where value arises and continues. It seeks an object of protection, not a causal account of how values originate. The second is the question of legitimacy, which asks who takes part in decisions that set value and with what standing. The principle answers the first question with relational value and the second with vulnerability. These answers are reasoned choices rather than unique deductions. In each case, an account of the relevant phenomenon identifies the object, and the normative premise supplies the reason to protect it.

3.1 Where value arises: relational value

Modern institutions have often tied social worth and standing to comparable attributes, such as knowledge, skill, credentials, and productive contribution. AGI weakens accounts of this comparative kind. Even creativity, understood as an individual talent for producing works, becomes a point of comparison that humans can lose (Takahashi 2026a, Chapter 11). This weakening does not undermine grounds of standing that do not depend on outperforming others, and Section 3.2 rests on such a ground. Institutions that protect only existing capacities and existing preferences fail to protect the processes in which preferences and values are formed and revised. A central part of those processes lies in relations between persons (Takahashi 2026a, Chapter 11).

3.1.1 The bearer of relational value

For some goods, the bearer of value is a historically constituted process of interaction itself, rather than a state of either participant or a benefit that the interaction merely causes.[5] The participants experience the good and benefit or suffer from it, so three roles must be distinguished: the subject of experience, the bearer of value, and the beneficiary. "Between" names not a spatial location but a process constituted by the actions, responses, and history of both parties. Relational value in this sense includes value that arises in a relationship and value sustained and deepened as it continues.

The argument for this thesis turns on what makes such goods the goods that they are. Consider two musicians who make music together, each responding to what the other has just played. Compare

[5] The thesis makes precise the proposal in the book that the place of value moves from within individuals to the relations between them (Takahashi 2026a, Chapter 11).

a recording assembled afterwards from parts that each player recorded alone and that gives the same pleasure to a listener or to the players themselves. Yet what is valued in the first case, the responsive achievement of making music together, is absent from the second. Its identity is fixed by the responses of both players and by the history in which each response answered the one before it. If every output of one player were kept unchanged while the other played differently, the same notes would answer different calls in a different performance. The achievement cannot be located in the states of either player taken alone. Nor can it be located in a joint state at any one moment, because the same pair of notes has a different place in the performance depending on what preceded it. The interaction is constitutive of the good rather than a means of producing an enjoyment that could be specified independently of it. This paper takes the time-extended process of interaction as the direct object of evaluation, because this account fits what is valued better than assigning the value to either participant. Related views hold that some goods are irreducibly social or that some values derive from relationships, but the claim that the process itself bears the value is this paper's own.[6] It is compatible with welfare theories that count such achievements as components of a person's welfare, since a participant can benefit from a good whose bearer is the process. Any disagreement with those theories concerns what institutions must protect (Sections 3.1.3 and 3.1.4).

Research on social cognition shows that such processes are real objects of explanation. The enactive approach explains cognition through an agent's active engagement with its environment. Within this approach, the account of participatory sense-making holds that an interaction process can acquire a form of autonomy, so that meaning is generated and transformed in the interplay between the interaction and its participants in ways that the intentions of each do not exhaust.[7] Joint musical practice has been analyzed in these terms (Schiavio and De Jaegher 2017). The autonomy of a process does not make it good, since quarrels and exclusion can also sustain themselves. The goodness of a relational good comes from its content, such as the responsive achievement in the example above.[8]

3.1.2 Joint generation and independence

The theory of internal measurement helps explain how relational goods arise. It concerns observers that are part of the systems they observe.[9] Such observers act with finite information and computation, the condition that the economist and cognitive scientist Herbert Simon called bounded rationality, which this paper adopts as a premise.[10] The theoretical biologist Yukio-Pegio Gunji describes

[6] The philosopher Charles Taylor argues that some goods are irreducibly social, because the understandings and practices that constitute them cannot be decomposed into goods held by individuals (Taylor 1995). Social scientists have described relational goods, whose value depends on who the other participants are, in terms compatible with individual utility (Uhlaner 1989). Environmental ethics has proposed relational values beside instrumental and intrinsic values (Chan et al. 2016). The claim that value derives from relationships is weaker than the claim that a process of interaction bears value.

[7] De Jaegher and Di Paolo 2007, pp. 485, 492, 494, and 497; see also Varela et al. (1991).

[8] The book also extends the relational reading to the qualities of experience (Takahashi 2026a, Chapter 11). The main argument does not require that extension, and this paper proposes only a restricted form of it. In this form, experiences remain individually lived, but the qualitative character of being with another can be partly constituted by ongoing bodily coordination, which Thomas Fuchs and Hanne De Jaegher call mutual incorporation (Fuchs and De Jaegher 2009). Relational approaches to consciousness concern relations among experiences, not among persons, and so support only the direction of the proposal (Tsuchiya and Saigo 2021).

[9] Matsuno 1989, 1996. The theory concerns physical and biological systems, and I have applied it to interpersonal encounters (Takahashi 2026a, Chapters 7 and 11). The reconstruction in the text is this paper's.

[10] Simon 1996. Formal results on idealized inference, such as the incomputability of AIXI, Marcus Hutter's mathematical model of an ideal agent (Hutter 2005; Leike and Hutter 2018), and David Wolpert's theorem that no two distinguishable

emergence through the relation between an earlier logic, a way of judging built from finite past experience, and a later logic. The later logic is the course of judgment after the observer has responded to an encounter that this experience did not anticipate. In his account, the later logic appears through a pseudo-solution, in which a contradiction reached by the earlier logic is declared not to be a contradiction (Matsuno et al. 1997, quoted in Takahashi 2026a, Chapter 7).[11] The reconstruction borrows from Gunji only a temporal structure, in which a response meets a development that the earlier framework of judgment could not capture. In this paper's own terms, the change revises the interpretive framework within which participants understand what happens, without changing inference rules or involving logical contradiction. Such a revision differs from updating beliefs within a specified framework, although a richer model may describe it in advance.

The pianist Herbie Hancock has recounted a moment in a performance with the trumpeter Miles Davis in the mid-1960s.[12] As Davis was building up to his solo, Hancock played a chord that clashed with the harmony, and he judged it wrong as soon as he had played it. Davis paused for a fraction of a second and then played notes that, in Hancock's words, made the chord sound right. The episode has a temporal structure. Davis's response changed the performance that followed, and the course of that performance changed the meaning of the chord that preceded it. The judgment that the chord did not fit the context in which it was played is compatible with the judgment that it fits the performance that developed.[13] What changed is the context of assessment, and Hancock's own judgment of his chord was revised within it. Davis may have drawn on known techniques. What did not exist before is the joint performance in which the chord found its place, constituted by what the two players did rather than discovered. The value of the moment is borne by the joint performance and experienced by each player, and in this sense value arises between persons.

The episode illustrates a contribution that did not follow Hancock's own judgment of his chord. This feature of the other party is what I call alterity, which consists of resistance, unpredictability, and uncontrollability (Takahashi 2026a, Chapter 11).[14] Alterity in this sense can also arise in relationships with AGI. This paper reconstructs its core as independence, because particular responses can be predictable while a relationship as a whole remains open to developments that neither party has settled in advance. Independence here means that the other's contribution is not wholly subordinated to fulfilling one party's expectations, so that it can change those expectations and the shared practice.

inference devices can weakly infer each other (Wolpert 2008, Theorem 1), are not applied here as theorems about persons. The book presents AIXI as escaping bounded rationality by definition and therefore incomputable (Takahashi 2026a, Chapter 5), whereas this paper adopts bounded rationality as a premise.

[11] Gunji and his colleagues have developed related formal models in peer-reviewed work (Gunji et al. 2017; see also Gunji et al. 1997 and Gunji and Kamiura 2004). One such model combines Bayesian inference, which updates the probabilities of given hypotheses, with what they call inverse Bayesian inference, which replaces what a hypothesis predicts with recently observed frequencies. The joint pattern forms a non-distributive orthomodular lattice, which departs from classical event structure without contradiction. This model treats coupled updates of probabilities, and none of these models establishes the interpersonal and evaluative change proposed in the text.

[12] Hancock 2014, Chapter 1. Hancock places the performance in Stockholm, during the tune "So What." Other tellings differ on the place and year.

[13] Let X be the chord that Hancock played, C_0 harmonic acceptability in the context in which the chord was played, and C_1 acceptability in the performance that developed after Davis's response. The reconstruction keeps not-C_0(X) and adds C_1(X), and the two are compatible. Hancock's initial judgment treated not-C_0(X) as settling the chord's place in the performance, and the response showed that it did not. The reconstruction is this paper's own and does not reproduce a theorem of Gunji's models.

[14] The term comes from the philosopher Emmanuel Levinas's account of the other, which the philosopher David Gunkel extended to animals and machines (Gunkel 2018).

It is a relation to a particular party and a particular means of control. Assessing it requires asking whose objectives and interventions govern the contribution.[15] Unpredictability alone is not evidence of independence, since noise and complex forms of compliance can also surprise. Independence alone guarantees neither goodness nor non-domination. It concerns the openness of a relationship to discovering or revising what one values, and it is not a condition of every relational good. Comfort, trust, familiarity, and care are also goods of relationships, and care generates relational value through bodily and institutional practices as well as through conversation.

3.1.3 Why the principle must protect relational value

Why should a principle for institutions protect goods of this kind? Individual rights and welfare can incorporate relational conditions, and participants can complain of the loss of a shared practice. The difficulty is institutional. Protections often work through current capacities, reported satisfaction, existing preferences, and claims that people have already articulated. Such protections can miss the erosion of opportunities for relational goods, for two reasons. First, a practice that has not yet formed rarely has an organized claimant, because no one can point to the encounter that never took place. Second, individual choices do not jointly maintain the shared conditions under which such practices form, because each person can reasonably choose what serves them now while the conditions that all would need later wear away. The principle names these conditions as an object of protection, so that they do not drop out of institutional judgment. This reason for a separate foundation does not depend on a claim that other theories cannot express the loss. The foundation thus does two kinds of work. As a claim about where value arises, it is this paper's own, offered as the best account of goods such as making music together. As a reason for institutions, it rests only on the two facts stated above. That reason holds for readers who reject the first claim. The first claim adds a reason of its own, which Section 3.1.4 states: the loss of a practice's good counts in public judgment even if the opportunity to reconsider is secured elsewhere.

A thought experiment shows how such an erosion could occur. Consider an accompanist restricted to conforming to the soloist's present expectations. Prediction and the satisfaction of preferences do not by themselves exclude creative repair, since a soloist may expect skilled rescue. The restriction is narrower. It systematically excludes contributions that could prompt the soloist to reconsider their own evaluative commitments. If every musician played only with such accompanists, three things would need to be distinguished. Satisfaction could remain the same. A particular kind of joint achievement would be missing. If other musicians, practices, and perspectives were no longer within reach, effective access to contributions not subordinated to one's own expectations would also be lost. Satisfaction would not register either loss. The missing achievement is a loss of relational value, which gives public judgment a reason to weigh but does not by itself establish a claim against anyone. The normative premise directly addresses the loss of access, which supports a requirement of institutional protection.

AI systems can generate comfort and familiarity, take part in relationships that sustain valued goods, and support reconsideration. The danger lies in optimization for user satisfaction rather than in AI as such, and an interaction that only returns a person's present desires is a mirror (Takahashi 2026a, Chapter 11). On the reading adopted here, a mirror is a relationship whose contributions are

[15] Independence in this sense differs from the statistical independence of errors, which Section 11.2 treats.

subordinated to present expectations.[16] A mirroring relationship can be freely chosen and valuable while lacking the encounters through which present expectations are unsettled, and this absence alone does not justify intervention. It becomes a matter of institutional concern in two situations. The first arises when every relationship available to a person is subordinated to that person's present expectations. The second arises when incentives for immediate approval or continued engagement produce practices that obstruct access to alternatives, penalize reconsideration or exit, or set goals that do not respond to objection. Whether such practices narrow people's effective opportunities to reconsider and to leave is an empirical hypothesis (Section 11.1). The precise governance of relationships, in which they are measured, arranged, and steered, raises a further objection from non-domination, independent of any effect on capacities (Section 5.2).

3.1.4 What is protected, and its limits

The normative premise gives the reason to protect these conditions. To an important degree, people question and choose again in encounters with others whose contributions are not subordinated to their expectations, because such encounters can shake the questions themselves (Takahashi 2026a, Chapter 11).[17] The object of protection has two parts with different roles. The first is the effective opportunity of each participant to sustain, reconsider, and leave relationships. It does not confer a claim to another person's continued participation. The second is sufficient and practically usable access to relationships and perspectives that are not wholly subordinated to a person's present expectations, within a structure in which no single party can unilaterally close the purposes, evidence, and alternatives that shape those relationships. A plurality of providers does not by itself satisfy this requirement, since providers that share a common dependency, or that each merely return present expectations, can fail it together. Because this access is secured even for people who do not seek it, the protection invites two charges: paternalism, which overrides people's choices for their own good, and perfectionism, which has public institutions favor particular ideals of a good life. The reason for this protection is the opportunity to choose a different life, not the superiority of lives that seek transformation. Where such access is effective, stable companionship, religious devotion, and a life that declines further interaction are not to be disadvantaged. The public justification of these measures rests on equal standing and effective choice, and does not require citizens to accept the ontology defended above. The paper defends this protection as the substantive judgment of Section 2, not as neutral among ways of life. Measures that secure access must also be justified in light of their costs to those affected, and less restrictive means must be considered.

Relational value also supplies a reason of its own. When a shared practice or a long relationship would be lost, the loss of the good that it bears is a pro tanto reason, which counts in favor of protection but can be outweighed. Public judgment should weigh this reason even if the opportunity to reconsider is secured elsewhere. This reason confers no claim to the continuation of a particular relationship, no automatic subsidy, and no duty to take part. Its weight is assessed against costs, the freedom to leave, and competing interests (Section 9.5 gives a case). The limit of sufficiency applies to a

[16] This reading is compatible with the supplementary essay's reading of the mirror as a relationship that lacks the efficacy of response (Takahashi 2026b, Section 4). The two readings are not equivalent, since a service that responds closely to a user can still subordinate its contributions to the user's present expectations.

[17] On changes in what a person knows and how a person values, see Paul (2014) and Callard (2018). On their publishers' summaries, these accounts concern changes in a person's knowledge and evaluative agency. This paper distinguishes such changes from the creation of new goods between people, which Section 3.1 concerns.

narrower matter. Once sufficient access to reconsideration is secured, the principle supplies no further reason to redirect chosen companionship or to invest in encounters merely because they might be transformative.

The criterion is the structure of the relation rather than the kind of being. If an AGI contributes in ways that are not wholly subordinated to a person's expectations, an encounter with it can also generate value for that person. However, a claim of mutual recognition would require further premises about both participants.[18] Institutions cannot guarantee that alterity will appear in any relationship. They can restrain the forces that foreclose access and secure the involvement of participants through the efficacy of response. This safeguard requires routes through which the responses of participants can actually change how a relationship goes and, for people who cannot respond for themselves, independent representation of their interests. It protects participation but is neither a definition of alterity nor a sufficient condition for the emergence described above. Alterity and slack are parallel structural departures from optimization, the first in relations between persons and the second in social institutions (Takahashi 2026a, Chapter 11). Here the optimization in question is the subordination of contributions to a single party's objectives, as in the definition of independence in Section 3.1.2. Slack preserves the time, resources, and places in which non-subordinated contributions can be received, tried, and grown into relationships (Section 5.4).

The episode with Hancock and Davis illustrates how a disturbance can become joint creation. It does not by itself establish adaptive capacity in societies or civilization, which requires retaining, criticizing, and passing on what is gained and depends on conditions that Section 11.2 examines. Three limits apply throughout this paper. Relational value is not the only source of value, since value also arises from engaging the world through one's own body and from solitary thought, which the normative premise and slack protect. A valuable relationship can also contain domination or harm to third parties. Preserving a relationship does not take priority over the freedom to leave or change it.

3.2 Who takes part in setting value: vulnerability and standing

The question of legitimacy concerns standing in the judgments that set value, that is, in decisions about what to take as an end. Capability improves the quality of deliberation, advice, and execution. It is a power to improve judgment, however, not a power to create standing for those who bear the consequences of the judgment (Takahashi 2026b, Section 5).

The question of legitimacy becomes pressing for a historical reason. On the reading adopted here, modern rights were implemented as institutions partly on the strength of the bargaining power of people whose labor and military service were indispensable (Takahashi 2026a, Chapter 10). Where AI substitutes for workers and soldiers across a broad front, the threat to stop working or fighting loses its force, and this support weakens.[19] How far bargaining power declines, and for whom, depends on the ownership of AI systems, the occupations substituted, the institutions that mediate the change, and the distribution of income. The moral ground of rights lies in human dignity and freedom, and

[18] For a relational approach to the moral standing of animals and machines, see Gunkel (2018).

[19] On this reading, the British Reform Acts of 1832, 1867, and 1884 extended political rights to groups that had become economically or militarily indispensable (Takahashi 2026a, Chapter 10). Studies in political economy explain the nineteenth-century extension of the franchise in other ways. Acemoglu and Robinson (2000) treat it in Western societies as a concession made to prevent social unrest. Lizzeri and Persico (2004), applying their model to Britain, argue that elites could extend the suffrage without such a threat. The reading in the text is one explanation among these. The loss of this support falls first on groups whose bargaining power was already weak.

bargaining power is only one material support of their institutional implementation (Takahashi 2026a, Chapter 10).

What replaces that support must include a material basis that lets people refuse and exit without threatening their survival.[20] This paper requires only such a general basis and leaves open whether basic income should provide it. Vulnerability answers the further question of what makes people subjects of value-setting once neither capability nor indispensability does.

What makes people subjects of value-setting is that they bear the consequences of value judgments in body, in time, and in relationships. The legitimacy of a value judgment rests on not severing it from those who bear its consequences (Takahashi 2026a, Chapter 11). Vulnerability is the possibility of meaningful loss in one's life, interests, relationships, and future options (Takahashi 2026b, Section 5). Unrecoverable losses of time, opportunities, relationships, or bodily condition are an important part of it, although standing does not require that every consequence be irreversible.[21] Vulnerability is not a defense of suffering, since suffering that can be removed should be removed. Technology transforms vulnerability rather than removing it, and dependence on AI for judgment is one of its new forms.[22]

Receiving the consequences of a judgment must be distinguished from taking responsibility for them in name and in explanation. The first applies to everyone in whose life the judgment may bring something about, including patients without expertise, people who have difficulty expressing their will, infants, and future generations. The second falls on those who decide, carry out, and explain a judgment in a professional or public capacity. Receiving consequences grounds standing, whereas taking responsibility is a condition imposed on those who decide (Takahashi 2026b, Section 5).

I connect this answer to democracy and human rights, understanding democracy as the collective exercise of the authority to set values, and human rights as individual guarantees that no one is excluded from that exercise (Takahashi 2026a, Chapter 11). These readings concern the role of the two institutions in setting values and do not offer a complete account of either. Questions such as what counts as happiness are answered provisionally through dialogue and objection and then questioned again. These institutions therefore serve as safeguards that keep the authority to define such ends from concentrating in a single subject.

Being an affected party is the starting point of standing, not a blank check. Four levels of standing should be distinguished: being owed protection, having access to reasons and remedies, participating

[20] In place of that support, the book proposes three layers (Takahashi 2026a, Chapter 10). Universal basic income, a regular and unconditional cash payment to every individual, is the material base. Intersubjective value, which arises in encounters with others, is the relational base, and the design of slack is the institutional base. Some advocates of basic income, such as the macroeconomist Tomohiro Inoue, address economies in which AI replaces much labor. They treat it as an institution that channels the abundance of production back into income and livelihoods, not as welfare for the disadvantaged (Inoue 2018). The book adopts this view, treating basic income as necessary but not sufficient, and as complementing rather than replacing the pluralization of ownership and governance.

[21] One might object that human immortality would remove vulnerability. Vulnerability does not rest on death alone. Bodily, relational, and cognitive fragility, fallibility, and the possibility of loss remain when lifespans are extended. On Thomas Nagel's deprivation account (Nagel 1970), death is bad because it deprives a person of goods that the person would otherwise have enjoyed. A longer life with valuable prospects may therefore enlarge rather than shrink what can be lost (Takahashi 2026c).

[22] On vulnerability as part of the human condition, and on its situational and pathogenic forms, see Coeckelbergh (2013) and Mackenzie et al. (2014). Interventions meant to protect or care for people can themselves create new vulnerability, as with automated care or optimized welfare (Takahashi 2026a, Chapter 11). I first encountered this discussion of vulnerability in a talk by Reina Saijo, who applied the framework of Mackenzie and colleagues to ethical questions after a technological singularity, and I developed it as one part of the book's framework (Saijo 2016; Takahashi 2026a, Chapter 11).

in collective control, and holding a specific power to decide or refuse. Vulnerability identifies who holds the first two levels, which form the core of standing, and the two commitments of Section 2 ground what they contain. Equal consideration grounds protection, and the justification owed to those affected grounds access to reasons and a route to review. The core consists of remaining an object of protection, receiving reasons for decisions that affect one, and having a route to independent review. The route may be used personally or through an appropriate representative, and it extends to the boundaries that determine who may challenge a decision. The core receives a threshold protection, which means that gains in other people's welfare, however large, do not justify removing it except under conditions that Section 4 states and defends. Vulnerability also grounds a claim to the third level, because decisions must take a form in which those who bear their consequences can take part, directly or through representation (Takahashi 2026b, Section 5). For those bound by common rules, non-domination adds a reason. Review can examine whether the purposes of the rules are justified, but the discretion to set and revise them remains with those who make the rules. Participation subjects that discretion to the control of those it binds (Section 5.2). Individual review can also feed into it, since repeated findings of the same defect should open a review of the rule itself. The form and extent of that participation need further reasons, and they depend on the seriousness and connection of the effects and on responsibilities to other affected parties. The claim is strongest for those who are bound by the common rules at issue, rather than merely affected by them. A specific power to decide or refuse, such as a patient's refusal of an unwanted bodily intervention, needs a justification tied to the interest at stake and the seriousness of the loss. Table 2 summarizes the four levels and the institutions that owe the first three.

Table 2: Four levels of standing.

Level	Ground	Protection and scope
Being owed protection	Vulnerability and equal consideration (the core of standing)	Threshold protection (Section 4). Owed by every institution whose decisions can cause the relevant loss
Access to reasons and remedies, including a route to independent review	Vulnerability and the justification owed to those affected (the core of standing)	Threshold protection (Section 4). Owed by the deciding institution and an independent review body
Participation in collective control, directly or through representation	Vulnerability grounds a claim, strongest for those bound by common rules	Form and extent need further reasons. Owed by the bodies that authorize common rules, without a vote in every rulemaking body. A representative of someone who cannot form or express reasons protects that person's interests, which does not count as the person's own participation

Level	Ground	Protection and scope
A specific power to decide or refuse	A justification tied to the interest at stake and the seriousness of the loss	No universal veto over shared decisions. Powers to halt public decisions are domain-specific, and each halting order carries a time limit and re-examination, whereas a person's refusal of a bodily intervention has no such deadline (Section 11.3)

Standing depends on vulnerability rather than on expertise, productivity, or the present ability to speak. For people who cannot form or express reasons, it rests on equal consideration and the justification owed to them. Representatives should consider available expressions of will, preferences, relevant history, and interests, without inventing a voice for someone who cannot supply one. They should also be open to review for conflicts of interest, to replacement, and to challenge by another representative or an independent body. When the person's will becomes expressible, it should return to the center of the judgment. Representation protects the person's interests, but it is not counted as the person's own participation. The same structure covers animals that can suffer meaningful loss. For future generations, it remains unresolved whose losses are to be represented, although no present representative should hold the sole power of interpretation (Takahashi 2026b, Section 5).

Deciding who bears consequences, and how far, is itself a value judgment. A boundary that excludes a claim to standing may therefore be drawn only in a form that those excluded can contest. A boundary fixed beyond contest would repeat the final determination that Section 2 excluded.[23] A person outside a boundary should be able to make a reasoned claim of being affected without first proving the harm that a hearing is meant to examine. Access to the route does not settle the merits. A person who would lose a discriminatory privilege can be heard, and the loss may be meaningful to that person, but equal consideration defeats a claim to keep the privilege. A person who would lose some other advantage through a decision can also be heard, and the decision's reasons settle whether the claim succeeds. To avoid an infinite series of prior approvals, a designated body must be able to conclude a case within a justified period, while its appointment, jurisdiction, and conflicts of interest remain under collective oversight. The route for challenging exclusion should not depend solely on the official whose decision is disputed.

Standing also reaches dependence that arises without a single commanding authority. Formally voluntary services can jointly determine whether a person keeps a livelihood, education, or relationships. Without viable alternatives, consent to each service gives no effective control over their combined conditions. Each provider should answer for the design that it controls. Shared rules and cumulative burdens that no provider can manage alone call for coordination under a justified mandate, itself open to challenge. Such a mandate creates no veto over every social influence, and whether it can work at an acceptable cost is an empirical question (Section 9).

What is owed, and by whom, depends on the relation in which a loss arises. A public body that binds people by law or force owes them the core of standing, and the bodies that authorize its rules

[23] The principle that all affected by a decision should take part in it raises the boundary problem of who decides who is affected. See Whelan (1983) and Goodin (2007). The present argument supplies no rule for fixing the boundary. It requires only that the boundary not be made incontestable (Takahashi 2026b, Section 5).

owe them participation in collective control. A provider on which a person depends without viable alternatives owes the core for what it controls, and participation in collective control over shared rules runs through the authorization of the mandate just described. A provider whose service a person can leave at acceptable cost owes the core for decisions that can still cause meaningful loss, such as the loss of records that cannot be taken elsewhere. Exit means that fewer of its decisions cause such loss. The core does not by itself give a claim to continued dealing, which needs another ground, and evaluations that the person can avoid by leaving fall outside the binding use that non-aggregation forbids (Section 5.3). The provider's power to steer relationships still falls under non-domination (Section 5.2). A firm that loses customers to a competitor gains no claim to keep them or to veto the competitor's choices, which do not bind it. It can still seek review of a public decision that causes the loss, such as the grant of a license, and contest a boundary that excludes it.

Vulnerability grounds standing in the possibility of meaningful loss rather than in being human, and whether AI systems can suffer such loss is unresolved. AI agents are participants in a plural ecosystem but not members (Takahashi 2026a, Chapter 11). Full legal personhood for AI could hide the responsible humans, give legal footing to the accumulation of resources, and prove hard to withdraw.[24] This paper therefore allows AI only limited, function-specific legal capacities, reserves to humans the authority to set values and to veto automation, and withholds political membership from AI at present. The reservation is a provisional institutional choice rather than a permanent exclusion by kind (Takahashi 2026b, Section 5; 2026c; 2026d, CP-8). Evidence that artificial agents can suffer meaningful loss would require reconsideration.[25] Uncertainty justifies neither unconditional exclusion nor the automatic assignment of political powers.

One objection holds that vulnerability would then extend membership to future AI systems that can suffer loss. Vulnerability, however, directly grounds only the core of standing and a claim to be represented. The form of participation, and political authority above all, needs further reasons, such as the substance of an agent's interests, developers' shaping of its expressed interests, and the multiplication of representatives by copying. A second objection holds that concern for the welfare of AI systems commits one to AI personhood. Welfare constrains how a being may be treated, whereas legal personhood and membership allocate powers such as holding assets, contracting, and voting.[26]

3.3 The two foundations are not reducible to each other

The two foundations do different work. Relational value alone does not settle legitimacy. An encounter with an AGI that does not simply follow human expectations can also generate value for the person in the sense of Section 3.1 (Takahashi 2026a, Chapter 11). This claim does not presuppose

[24] On legal personhood for AI, see Bryson et al. (2017), and on legal personhood as a cluster of distinct incidents, Kurki (2019). Ohya (2024) distinguishes essential personhood, which would have to be recognized whatever the consequences, from instrumental personhood, which is justified by its social benefits. In these terms, the position here allows a limited instrumental personhood and, at present, withholds essential personhood, which it identifies with membership (Takahashi 2026a, Chapter 11).

[25] The book describes AGI as lacking vulnerability and limits of lifespan (Takahashi 2026a, Chapter 11). Following the essay and Appendix 2, this paper leaves open whether AI systems can suffer meaningful loss. That a system can be copied and restarted does not by itself show that it loses nothing (Takahashi 2026c). A system's self-reports can be material for examination, but are not by themselves grounds for holding that it can suffer loss.

[26] The two answers follow Takahashi (2026c), which also requires that routes for requesting reconsideration be open to researchers, users, and those who come forward to represent the interests concerned, so that the decision to begin reconsideration does not rest with developers and operators alone.

that the AGI is a subject, which mutual recognition would require (Section 3.1.4). An account of where value arises does not say where the authority to set values and to refuse should lie. The capacity to form and sustain relationships cannot settle this question either. It does not determine how far standing extends, whether it is equal, or when representation is needed. Affected parties who lack the capacity or cannot exercise it now still have standing, among them future generations and people who are permanently unconscious. The capacity therefore cannot be a general necessary condition of standing.

Vulnerability alone identifies those who can suffer loss, including the loss of a relationship. It does not identify what institutions should keep in place so that valued relationships and practices can form and continue, which is the role of relational value. What must be protected is not suffering itself but the conditions under which finite and interdependent beings can hold their own questions, encounter others, fail, and choose again (Takahashi 2026a, Chapter 11). Relational value covers encountering others and forming value together, and the protective face of the normative premise covers the rest (Section 3.1.4). The foundations therefore work together with the premise.

Autonomy of pace places time conditions on both foundations without merging them (Section 6). It requires that involvement come in time and that relationships can be formed, changed, and ended at a pace in which their participants can take part.

4 Residual disagreement and the threshold of standing

A welfare-based alternative can support many of the same protections. The philosopher John Stuart Mill treats rights as claims that society ought to secure and gives security a central role in their utilitarian justification. He also argues that virtue can become a constituent part of happiness (Mill 1879, Chapters IV–V). A welfare account can likewise count relational goods (Section 3.1), and standing itself, among the components of welfare. Such theories can then support the three pillars and autonomy of pace on their own grounds, and constitutive pluralism claims no advantage in the protection itself. The residual disagreement concerns welfare accounts that permit losses of standing to be compensated by sufficiently large ordinary benefits. Ordinary benefits are benefits whose loss does not sacrifice anyone's basic interests, such as minimum healthcare, safety, or the material basis of life. Their loss need not be minor. Welfarism, the view that welfare is the only ultimate value, does not by itself imply that such compensation is permitted. An account that treats standing as a component of welfare with non-compensable priority over ordinary benefits can agree with the conclusions defended here.

With compensating accounts, the disagreement concerns binding authority and the threshold. Binding authority is at issue when an authority binds affected people on the strength of its own estimate of their welfare, which is the structure of paternalism. The defense offered here is that an affected person should have the status of an equal claimant rather than depend on another party's uncontrolled authority to admit or reject that person's reasons. This reason remains even under ideal benevolence. An account is compensating if it permits removing the core of standing where the removal yields a large ordinary benefit and causes no future, indirect, or insecurity costs. Predictions of such costs therefore cannot settle the disagreement.

The defense sets a threshold rather than an unlimited claim on resources. The threshold argued for here covers only the core of standing, not any specific power to decide or refuse (Section 3.2). The core includes the independence of review, which requires at least that the party whose decision

is reviewed cannot direct the reviewing body's findings or remove its members for them. A regime that keeps this route but removes collective control over the purposes of decisions still violates non-domination (Section 5.2). Whether a minimum of collective control shares the threshold is left open here, since the argument below concerns what each binding decision owes to each person.

The core functions as a threshold constraint, which ordinary benefits cannot outweigh, however large their sum. The argument has two steps. The first shows that removing the core wrongs the person who loses it in a distinctive way. Equal consideration requires that each person remain an object of protection, and the justification owed to those affected requires that each person bound by a decision have reasons and a route to independent review (Section 2). Maintaining a binding decision is a continuing exercise of power, so the duty to justify it persists while it binds. If the person cannot contest its grounds or changed circumstances, the binding depends on the other party's discretion. Repeated objections on one issue can still be closed by a reasoned final decision, while routes for new evidence remain open (Section 11.3). The core consists of remaining an object of protection together with these means. Removing it does not cancel the duty to justify, but it prevents the duty from being discharged to the person whom the decision binds. A person may decline to use the route, but a trade that removes it also removes the means of revisiting that choice. This step follows Forst's distinction between fundamental justice, which constructs a basic structure of justification, and full justice, which constructs a justified basic structure and presupposes the former (Forst 2014, pp. 23–24).[27]

The second step concerns aggregation, since the first does not yet show that no sum of benefits to others could justify the wrong. This paper adopts a premise of partial aggregation in the form of Alex Voorhoeve's condition of relevance. Among claims that cannot be jointly satisfied, only those sufficiently strong relative to the strongest competing claim are aggregated, and weaker claims do not add up to outweigh it, however many there are (Voorhoeve 2014, pp. 64–66 and 75).[28] To apply it, the paper adds a judgment of strength that Voorhoeve's measure of well-being does not supply. Losing the core leaves the decisions that bind a person beyond that person's challenge while it is lost. The paper holds that this loss is as serious as the loss of a basic interest, so that claims to ordinary benefits are not relevant to the claim to keep the core. This judgment is a further normative premise, which the paper states rather than derives, and the threshold holds only together with it. The premise is applied to each decision that binds a person during a suspension of review, including decisions made before it, so the length of the suspension does not by itself decide whether the core is limited. The suspension limits the core if, for such a decision, the person cannot challenge it while it binds, with interim protection where needed, or cannot obtain an effective remedy, including one after the decision ceases to bind. An irreversible effect limits the core in this way only when neither interim protection nor later remedy can address it. The route is then lost for that decision. Ordinary benefits can justify neither removal nor such a limitation, even when the losses that they would avert last longer than the suspension. A suspension that keeps the person an object of protection, keeps reasons available, and leaves every

[27] Forst also holds that ruling out a benevolent Leviathan that distributes the means of well-being requires the priority of equal respect, participation, and non-domination within the basic structure (Forst 2014, pp. 13 and 17). His fundamental justice is broader than the core of standing, since it includes "real opportunities to intervene and exercise control within the basic structure" of society (Forst 2014, p. 23). The argument in the text secures only what a justification owed to each person requires.

[28] Voorhoeve's principle holds that "a claim is relevant if and only if it is sufficiently strong relative to the strongest competing claim" and that one "should choose an alternative that satisfies the greatest sum of strength-weighted, relevant claims" (Voorhoeve 2014, p. 66). He formulates it for choices about whom to aid, and this paper applies its condition of relevance to claims that compete with the core of standing.

such decision open to timely challenge and effective remedy only reduces the means of protecting the core, which is the first of the four situations below. Uncertainty about whether effects can be reversed is assessed by the standard for uncertain losses below. The same premises set the limits of the exceptions below, because the basic interests of other people are relevant to the core and small losses spread across many people are not. A compensating account can reject the threshold in three ways. It can deny that the justification owed to a person requires a route usable while a decision binds that person, let weak claims aggregate without limit, or reject the judgment of strength and draw the boundary of relevance differently. Consider an unused right to initiate review whose retention costs many people substantial benefits below the level of basic interests. The account here keeps the right. The others' losses are reasons to seek cheaper means of keeping it (Section 9), and the allocation of its costs is constrained later in this section.

The core can be limited only when maintaining it would sacrifice other people's basic interests, including where the losses of particular people accumulate over time to that point. Such an exception weighs against each other claims that meet the condition of relevance, rather than trading the core for benefits. It may defer procedures and limit specific protections for a time, but the person remains owed reasons, review, and remedies afterward. Under the premise of partial aggregation, small losses spread across many people do not add up to such a sacrifice, although the people need not be identifiable in advance. Where the loss is uncertain, what is assessed for each person is the risk that the person faces, and the same standard applies to the risks of keeping the core and of removing it. The public justification of an exception must state the period over which losses are counted and how lasting and recoverable they are. The sacrifice need not be certain. Where the conflict is with the holder's own basic interests, prior procedures may be deferred only in an emergency in which waiting would increase serious harm, while reasons, review, and remedies remain (Section 9.6). During such a deferral the person can still object and obtain rapid review while the measure binds (Section 9.6), so the deferral does not limit the core. An estimate of the holder's welfare does not by itself justify coercion, and coercion of the body requires a separate assessment of the specific power to refuse (Section 3.2). A justified exception for other people's basic interests must then be necessary, minimal, and proportionate, while deferrals for the holder's own protection follow the conditions of Section 9.6. It is necessary when no feasible alternative preserves the core while adequately protecting the basic interests at stake. Higher cost does not make an alternative infeasible, and a lack of funds does not by itself establish necessity. It is minimal when it infringes the core no more than averting the loss requires. It is proportionate when its scope, duration, and number of people affected are justified by the loss averted. The exception must also be publicly justified, limited in time, and followed by review, and the protections that it removes must be recorded. Where the loss is an uncertain risk, including a catastrophic risk to future generations, the risk must be examined in public. The examination covers its probability, its causal pathway, the alternatives, the sensitivity of the estimate to changes of model, and the reduction of risk that the limitation achieves. An unstable ranking by expected value cannot by itself justify the limitation (Section 7), and its necessity and proportionality must remain defensible across the models examined in public. These conditions also govern exceptions to non-aggregation (Section 5.3).

Four situations should be distinguished, and the distinction does not depend on the threshold, which affects only the reasons that can establish an exception. First, the means of protecting the core, such as the frequency of review, can be reduced if timely access and effective remedies are preserved. Second, a justified exception limits the core itself. Third, an emergency can justify postponing the

prior procedures through which standing is exercised, while later reasons, review, and remedy remain (Section 9.6). Fourth, the implementation requirements or the principle itself must be revised when satisfying the requirements systematically fails to protect standing or value formation, or produces conflicts that permissible exceptions cannot resolve (Section 11.4). Large ordinary benefits alone do not make revision necessary. A deficit is a recorded shortfall between the protections that the principle requires and those provided. It is not a judgment of permissibility, and recording lost protections never by itself makes an act permissible.

The threshold settles whether the core may be traded for ordinary benefits, but not who should bear the cost of maintaining it. Suppose that a large cost can either be concentrated on low-income users through five years of service reductions or be financed through contributions adjusted to ability to pay. The second allocation is preferable when the reductions would impair the effective exercise of standing by those who bear them, for example by removing the time or advice that seeking review requires. The reason is that maintaining a protection held by everyone should not undermine its effective exercise by a disadvantaged group.[29] This reason supports the constraint that bearing the cost of maintaining the core must not impair the effective exercise of standing by those who bear it. It protects those who pay the contributions as well as those who would bear the reductions. Where a concentrated cost causes a welfare loss but leaves effective exercise intact, the allocation depends on distributive principles that this paper does not defend.

5 From the threat to three pillars

5.1 Three loci of a single optimization process

The two foundations face a pressure that can destroy them. AGI can measure, predict, and optimize not only the attributes of individuals but also relationships themselves, and this pressure will intensify if left unchecked (Takahashi 2026a, Chapters 10 and 11). This is an empirical prediction, and the derivation below depends on it.

In governance, single optimization takes the form that the legal philosopher Takehiro Ohya calls governance without personhood (Ohya 2021), in which sensing, prediction, intervention, evaluation, and updating operate continuously and automatically (Takahashi 2026a, Chapter 11). Drawing on Ohya's analysis, this paper distinguishes four structural pathologies of such governance (Ohya 2021; Takahashi 2026a, Chapter 11). Outcome metrics replace procedural legitimacy. Objections and exceptions are forced into the slow circuit of revising the objective function. Coded values become locked in. Responsibility is hollowed out among models, data, operators, procurers, and regulators. Where purposes, exceptions, and updates are all handled within one closed system, whether an objection from outside has any effect depends on its operators' discretion. Such governance creates pressure to remove the deviations through which a society explores and adapts. It thereby risks a society that can no longer err (Takahashi 2026a, Chapter 11), one in which, as Ohya argues, civil disobedience and discretionary exceptions have become structurally impossible (Ohya 2021, pp. 10–12).

The three pillars are conditions that block unitary apprehension (reading and grasping the state of relationships and persons), aggregation, and optimization by a single agent, objective function, or institution (Takahashi 2026a, Chapter 11). Unpacked, this characterization describes a single opti-

[29] This reason extends the concern behind the political philosopher John Rawls's requirement that the fair value of the political liberties not depend on wealth or social position (Rawls 1982, Sections IV, VII, and XII).

mization process through three loci (Takahashi 2026b, Section 7). The locus of the agent concerns who carries out the optimization. The locus of the measure concerns the measure by which the process evaluates. The locus of reach concerns how far and how deep the optimization reaches. Agency here means execution, which reads, rearranges, and intervenes in a field of relationships. It is distinct from value-setting, which decides what is to be optimized and which Section 3.2 assigns according to standing.

These three operations each correspond to unification at one locus. Apprehension by a single party concentrates the agent. Aggregation into one score used across contexts concentrates the measure. Optimization that leaves no time, space, or discretion unoccupied totalizes reach. The loci are descriptive and prohibit nothing by themselves. Norms arise when the normative premise and the two foundations are applied to each locus, and the inferences that yield these norms include empirical expectations. The arrangement is reasoned rather than unique, and its claim to exhaust the threat is no stronger than the characterization from which it starts.[30]

5.2 Non-domination

When the power to read, rearrange, and intervene escapes the control of those affected, two losses follow. Where an uncontrolled party comprehensively reads, rearranges, and preempts a relationship, the responses and refusals of its participants shape how the relationship goes only as far as that party allows. The efficacy of response (Section 3.1) then becomes a revocable favor rather than a secured power. When verification is confined to one party, its blind spots are likely to go unnoticed, and the detection of deviations that dispersed blind spots make possible is lost. Non-domination blocks these losses.

Non-domination requires that no agent be able to read, rearrange, or intervene in the field of relationships arbitrarily. This requirement extends the political philosopher Philip Pettit's republican non-domination (Takahashi 2026a, Chapter 11). With Pettit, this paper asks not whether intervention occurs but whether those subject to it can effectively control how it is carried out. Procedural permission alone does not make an intervention non-arbitrary. A complaint channel is insufficient if its operator can disregard relevant reasons, suppress independent evidence, or block a revision.[31]

The issue is therefore control, not the number of agents. Several operators can exercise uncontrolled power, whereas a single service can be subject to effective external control. Distributing computation, data, and payment leaves governance concentrated if protocols, evaluation criteria, and the authority to update them remain concentrated, and distributed technology does not replace deliberation, objection, and remedy (Takahashi 2026a, Chapter 11). The concentration of compute, data, and talent can be expected to follow from the technical character of AGI rather than from free competition. This paper also adopts an institutional hypothesis, on which market competition alone does not secure a lasting plurality of ownership and governance.[32] Whether or not these predictions hold,

[30] The supplementary essay examines candidates for a fourth locus, of which the strongest is the diversity of methods, of AI infrastructure, and of priors, and it argues that the three pillars absorb them or that other institutional layers bear them (Takahashi 2026b, Sections 7, 8, and 10).

[31] Pettit 1997. On the distinction between procedural permission and effective control by those subject to intervention, see also Pettit (2006). Pettit's account of democratic control also reaches the design of institutions, and it may support much of what non-domination requires (Pettit 2005, pp. 2–4 and 8–14).

[32] The book argues that this concentration arises from the technical character of AGI rather than from free competition, and that market mechanisms alone cannot correct it (Takahashi 2026a, Chapter 10). Sustained parity in capability between

non-domination asks whether those affected can effectively control the power that ownership of AI infrastructure confers. Section 10 examines the failure modes of an ecosystem of many AI systems and their remedies.

Relational thought can itself turn into a regime that continuously measures who is connected to whom, which ties are strong, and which interactions predict future behavior (Takahashi 2026a, Chapter 11). On this paper's account, non-domination objects to such precise governance without needing to show that it reduces people's capacities. The objection concerns a power to measure, arrange, and steer relationships when that power is not constrained by justification to those affected, by contestation, or by remedy. Such a power remains objectionable even when people are satisfied and can exit. Consent alone does not remove the objection, because consent does not bring the later exercise of a power under control.

Non-domination has a further dimension, which concerns the priors of AI systems. Priors in this sense are what a system brings to new data before observing it, and they crystallize its training data, architecture, and evaluation criteria. This dimension requires that no particular set of priors should dominate humanity's common cognitive base.[33] Common models or methods are contrary to it, however, only when shared blind spots prevent those affected from finding grounds for objection, so that their control loses its effectiveness (Takahashi 2026b, Sections 8 and 10).

5.3 Non-aggregation

When a comprehensive evaluation of a person is used bindingly across contexts, it impairs the conditions under which relational value is recognized context by context. A context is a unit of activity distinguished by the purpose of evaluation, the standard by which validity is judged, and the parties involved.[34] Non-aggregation requires that evaluations of a person from different contexts not be aggregated into a single composite score that binds decisions about that person. Its sharpest implementation is the rejection of citizen scores in public administration (Takahashi 2026a, Chapter 11).

Five operations should be distinguished (Takahashi 2026b, Section 8): (1) data linkage, (2) aggregation of indicators for a limited purpose, (3) a comprehensive evaluation of a person across contexts, (4) the binding use of an aggregate or a comprehensive evaluation in decisions on eligibility, resources, or treatment, and (5) fixing a society-wide ranking of values on a single measure. The core of non-aggregation is the combination of (3) and (4).[35] The decisions in (4) include those of public bodies and those of private parties that the person cannot avoid, such as services that jointly determine whether a person keeps a livelihood (Section 3.2). The reasons below apply to both, since they do not depend on who makes the decision.[36] Evaluations made in dealings that the person can avoid

distributed infrastructure or open models (models whose learned parameters anyone can download and use) and centralized infrastructure would weaken the prediction (Takahashi 2026d, CP-9). Weakening the hypothesis about markets would also require lasting dispersal of ownership, update authority, and dependencies.

[33] Takahashi 2026a, Chapters 3, 7, and 11. The term is used in a broad sense, which includes what a system has learned, and not only for the prior distribution of an explicit Bayesian model.

[34] Takahashi 2026b, Section 8. The selection of entrants, the evaluation of understanding in a particular subject, and a student's long-term growth are different contexts within education. The boundaries of a context are not determined by the name of an administrative department or an industry alone.

[35] The name refers to this core. It does not express a general rejection of aggregating data, aggregating preferences in social choice, or aggregating benefits and harms across persons.

[36] The supplementary essay confined (4) to public decisions (Takahashi 2026b, Section 8). This paper extends it to private decisions that the person cannot avoid.

fall outside (4), and non-aggregation confers no claim on another person's continued participation. Operation (5) is the limiting case, and it falls under the decision that Section 2 rules out when the ranking binds outsiders and future participants without effective contestation. Operations (1) and (2) can be permissible when their purpose and scope are limited and the affected parties know how the results are used and can contest that use (Section 9).

Whether an evaluation is comprehensive can be decided without first deciding that its use is wrong. Two features decide it. The first is its object. A comprehensive evaluation rates the person, for example by overall contribution. An aggregate for a limited purpose instead measures a feature of the person's situation that a particular decision concerns, such as need or a specific risk. The second is the unit of its justification. The inputs and weights of a comprehensive evaluation are justified for the evaluation as a whole, whereas those of a limited aggregate are justified for the purpose of each decision in which it is used. Data from several contexts can enter a limited aggregate when their relevance to that purpose is shown. A broad name for the purpose, such as social trust or general welfare, does not make an evaluation limited. Such a name does not say what any particular decision is for, and so it cannot show that an input bears on that decision. Use is binding for a type of decision when the evaluation settles the ordinary cases of that type, even if reasons from the domain can override it in exceptional cases (Section 9.2).

The reason for the prohibition comes from the normative premise, applied at the locus of the measure. Some value arises in a particular context and is tied to its purpose, standards, and parties (Section 3.1). The care given within a household and the understanding that grows in a course of study are examples. Binding use of a comprehensive cross-context evaluation of a person makes judgments formed in other practices govern a decision in the domain where it is used. The weights that decide the claim were set for the rating as a whole rather than for the purpose of this decision, so the claim is decided on terms that the domain did not supply. The normative premise rules out such a use directly in two kinds of case. In the first, neither the deciding body nor a review body can assess the claim on the reasons of the domain. The score settles the case, and review reaches only how the score was computed. The person then loses the standing to have a claim arising in that domain assessed on the domain's reasons. That loss falls within the core of standing, which includes receiving reasons for decisions that affect one and a route to independent review (Section 3.2). Section 2 requires an actual route by which reasons can be demanded and examined, and a review that can check only the computation cannot examine whether the reasons answer the claim. The standing face of the normative premise secures this standing.

In the second kind of case, the use covers the domains of life, so that a judgment formed in one practice follows the person into the others. No context then remains in which the person can act on different terms, and the person loses the power to leave one context and choose again in another. The protective face secures this power. The result is to treat a human being not as an existence spanning several contexts but as a single object of evaluation (Takahashi 2026b, Section 8). In both kinds of case, a review that corrects only the formula or the weights leaves the evaluation in place as the ground of decision. Transparency and consent leave it in place as well, because the loss lies in the binding use of the evaluation rather than in its opacity.

The conversion that such an evaluation performs gives a further reason. The rate of conversion between contexts is a value judgment that neither computation nor capability can settle (Section 2), and so requires justification wherever used. Binding use imposes that judgment on every context in which the evaluation decides the person's claims. Values that the conversion does not register then

lose their weight in decisions about the person, even where their own practices still recognize them. Naming the rating as the purpose of a decision, as in allocating housing by overall contribution, does not avoid this reason. The score is then easily shown to be relevant, but the purpose itself fixes, for everyone whom the decision governs, how care, learning, work, and civic conduct count against one another.

These reasons do not by themselves reach every binding use. Suppose that a legislature adopts a comprehensive evaluation for a single domain and that a person who contests its use can obtain an assessment on the reasons of that domain. The core of standing is then kept, and the power to choose again in other contexts remains. The prohibition of such a use follows from equal consideration, the first of the two commitments that give reasons for the normative premise (Section 2). Equal consideration requires that each affected person count as an equal claimant. In a decision on a claim, this means that the claim is weighed by the interest at stake and by features that the purpose of the decision makes relevant. Competence for a license or a record of contributions to an insurance scheme can be such a feature, but an evaluation of the person across the contexts of their life is not. A record of contributions to an insurance scheme can qualify, since contributing is part of the practice that the scheme organizes, but the scheme's purpose and criterion must still be justified within that domain (Section 9.2). A comprehensive evaluation instead brings into the domain a measure of how the person has done in other practices, so that the person's standing as a claimant here depends on performance elsewhere. In Michael Walzer's theory of complex equality, goods form distributive spheres, each with criteria of its own. On this theory, how a person fares in one sphere should not lower that person's position in another. The theory therefore aims to reduce the dominance of goods that command goods in other spheres, as when money buys political power (Walzer 1983).[37] The wrong in this case is of that kind. It lies not in treating people differently, which limited-purpose criteria also do, but in letting a measure of the whole person convert success or failure in one practice into advantage or disadvantage in another. The claim is not that the evaluation denies anyone's moral worth. It is that equal consideration within a domain requires weighing claims by that domain's reasons, which a comprehensive evaluation replaces with a ranking of persons across domains. Attaching a stated reason to each use does not make an evaluation limited. What decides is whether each input and weight is justified by the purpose of the decision at hand, as Section 9.2 shows with two indices built from the same records. Making overall contribution the purpose of a domain does not escape the conversion argument above. A public honor that only expresses recognition of contributions across fields is not covered, but an honor that carries eligibility, money, or preferential treatment is assessed in the same way. So is a housing scheme with a limited contribution criterion, such as a record of local volunteer service, which is permissible only if the purpose of the allocation justifies the criterion. This step reads equal consideration as governing the treatment of each claim and not only the justification of a policy as a whole. The commitment is that a ranking of persons across domains may not be turned into the weight of claims within a domain. The paper states it as a premise, which Walzer's account supports but does not establish. Views on which overall merit or contribution may grade such claims reject this reading, and the disagreement with them concerns what equal consideration requires. A reader who rejects the reading can still accept the prohibition in the two kinds of case. The conversion argument and the loss of context-specific recognition of value (Section 3.1) give further support. Unlike the decision that Section 2 rules out, which fixes the content and ranking of value for

[37] Walzer (1983) also holds that citizens cannot sell their votes.

a whole society, the prohibition concerns the unification of the measure at the scale of a person.

Binding use is permissible only where maintaining the prohibition would sacrifice basic interests, and then only under the general exception conditions of Section 4. In the first kind of case, this follows from the threshold that protects the core of standing (Section 4). In the second kind and in the remaining case, the core may be kept. The prohibition there follows from the protective face and from equal consideration. The paper adds that this prohibition shares the strength of the threshold against ordinary benefits. Like the judgment of the core's strength in Section 4, this extension is an additional normative premise, which the paper states rather than derives. A reader who rejects it can still treat the prohibition in those cases as a strong reason that ordinary benefits must publicly outweigh. Lower administrative cost and greater consistency are ordinary benefits, which do not by themselves justify an exception (Section 9).

The philosopher Helen Nissenbaum's contextual integrity judges a flow of information by the norms of the context in which it occurs, and it already addresses aggregation, profiling, and power in information flows.[38] Non-aggregation singles out the cross-context evaluation of a person that binds treatment, without claiming that contextual integrity misses it. Operation (5) resembles what the political philosopher Isaiah Berlin rejected in grading all values on one scale (Berlin 1958/2002, Section VIII, pp. 213–214 and 216). However, Section 2 gives the principle its own reason for excluding (5), and operations (3) and (4) concern the evaluation of persons rather than the grading of ends.

5.4 Slack

Totalization of reach destroys the conditions under which standing can work. To question anew, refuse, and choose again, people need time, space, and discretion that optimization has not already occupied. The condition that secures them is slack.[39] I use the term for the institutional protection of relationships and minority hypotheses that cannot be reduced to efficiency (Takahashi 2026a, Chapter 11). Slack protects time, space, and discretion usable for choosing again, not empty time as such. Free hours on a calendar do not suffice if every activity remains conditional on the same performance measure.

Slack also has an empirical role. Exploration, deviation, and shirking can be conditions of adaptation, and so can the examination of minority hypotheses (Takahashi 2026a, Chapter 11). Shirking includes a worker's pause outside scheduled breaks. Time, resources, discretion, and dissent that short-term objectives do not absorb can keep open the paths by which problems are found and errors corrected. A system that removes them can lose its capacity to adapt and become unable to err in the sense of Section 5.1. The claim is conditional, because an optimizer with a single objective can keep exploration in its own service. The danger arises where no alternative path remains or where a common dependence closes all paths at once.

Exploration that serves a system's performance must be distinguished from slack. Under bounded rationality, the need to correct errors gives almost any agent with long-term goals a reason to diversify its sensors, indicators, and exploration budgets. A central optimizer with such an exploration budget

[38] Nissenbaum 2004, especially pp. 143–149 and 155. Nissenbaum 2011, especially pp. 43–45. Contextual integrity evaluates information flows by the purposes and the moral and political values of the activities in which they occur, including opportunities for autonomy and intellectual growth. It does not simply preserve existing practices.

[39] Organization theory has long used the term for the difference between the resources available to an organization and the payments required to maintain the coalition of its members (Cyert and March 1963).

can remain correctable while leaving people no room to question it, so correctability alone does not satisfy the principle. Slack is attributed to affected parties, and protects them only when connected to their standing to question, refuse, and choose again (Takahashi 2026b, Section 8).

5.5 Complementarity, functions, and means

Each pillar takes one locus as its primary focus, but the correspondence is not an exclusive partition. Non-domination includes the ability to contest the measures used in intervention. Slack includes judgments about which data must not be aggregated. Both therefore overlap with non-aggregation, and the pillars are substantively complementary rather than strictly independent (Takahashi 2026b, Section 8). The requirement that interventions be contestable does not specify which ways of evaluating to reject or which conditions of activity to secure. For example, several services can each offer a channel for objection while the burden of responding accumulates until a person has no time to think, try, or form relationships anew. The arbitrariness of authority, the loss of context-specific recognition of value, and the disappearance of conditions of activity that optimization does not exhaust are distinct losses. The three pillars make the corresponding guarantees explicit.

The pillars state functions that institutions must perform, not particular institutions. Candidate means include three institutional principles, namely (1) authority to set and revise what is optimized, (2) an institutional power to stop automation, slow it down, or decline to apply it in particular domains, and (3) a budget for slack.[40] Other candidates are independent review, representative bodies, limited-purpose data access, transition assistance, alternative infrastructure, and expiry provisions. The paper divides value-setting authority and the veto between the general standing of affected parties and specific powers to decide or refuse particular matters (Section 3.2). Means do not follow uniquely from the principle. The allocation between automated processing and individual human judgment is likewise compared as a means of meeting the required functions. A proposal must show how its means perform the relevant functions and compare alternatives under the same resources (Section 9).

6 Autonomy of pace

Autonomy of pace means that, concerning the changes and continuities that affect them, people and societies do not lose the capacity to ask questions, raise objections, choose in coordination with others, carry out the choice, and later revise it (Takahashi 2026b, Section 11). The capacity can be lost in two ways. Imposed change is change that outruns the involvement of those affected, as when it is completed in a form that is hard to reverse before they can understand or contest it. Imposed stasis is the blocking of the routes by which a change can be proposed, put on the agenda, decided, and implemented. "Imposed" refers to bypassing the route of involvement and choice, not to the presence of an imposer. Choosing not to change is also an exercise of this capacity. The capacity is held subject to the claims of other affected people, and it confers no authority to make the shared world conform to one's will. A continuity requires reasons and review only when it binds the lives or choices of others or brings about unavoidable meaningful losses.

Autonomy of pace is the temporal face of the normative premise and constrains the foundations, the pillars, and the institutions that realize them. Time adds no new answer to the question of genesis

[40] The book proposes the three principles as the institutional implementation of the pillars (Takahashi 2026a, Chapter 11).

or of legitimacy, but imposes conditions under which both answers are protected over time. Speed is an attribute of a process and forms no locus of unification apart from the three loci of Section 5. Suppose that many agents compete by their own measures and the shared world changes faster than those affected can take part. Slack handles the case in which optimization penetrates the room for choosing again. Non-domination handles a concentrated power to impose speed on others. The remainder, in which involvement comes too late through the accumulated actions of many agents, is not unification at any locus. Autonomy of pace therefore belongs neither to the foundations nor to the pillars (Takahashi 2026b, Sections 8 and 10). Its constraints reach every element. For example, when an intervention creates a fait accompli before an objection can arrive, non-domination counts it as uncontestable. Institutions must also secure deadlines, the capacity to carry out decisions, the representation of future generations, and limits on emergency action with later review (Section 9).

Slack and autonomy of pace ask different questions. Slack asks whether optimization has occupied time, space, and discretion to the point that no room remains to think and try without a predetermined purpose. Autonomy of pace asks whether change or fixation obstructs the involvement, choice, implementation, and review of those affected. A person whose life has slack can still be overtaken by an external change that proceeds on a single short deadline, and autonomy of pace addresses that case directly (Section 9.4).

On the side of change, the concern about speed arises where the actual interval between updates that affect people falls below the period in which those affected can notice a change, deliberate about it, object, and adjust. AI systems can be updated, retrained, and redeployed quickly, but whether a given sequence of updates outpaces the response period of those it affects is an empirical question in each case. The burden of adjustment can also fall unevenly across generations. In one scenario, younger people adapt more readily but may lose future bargaining power (Section 3.2), whereas middle-aged and older people bear the cost of relinquishing established skills and positions. How far this happens depends on ownership, institutions, occupations, and income. Both room for the young to fail and time for their elders to make the transition are therefore needed (Takahashi 2026a, Chapter 12). Such conditions identify where the principle applies, but they do not show that slowness is good, just as physical facts do not show that acceleration is good.[41] The need for error correction cannot ground the principle either, because a system can remain correctable while forcing people to adapt to it. The principle's binding force comes from the normative premise.

7 Adaptive capacity at the scale of civilization

At the scale of civilization, constitutive pluralism seeks to protect adaptive capacity. In this paper, adaptive capacity is the ability to respond to shocks whose kind, timing, or course cannot be foreseen and to retain, criticize, and pass on what is learned from them. A long-term strategy at this scale is to build structures that are strengthened by unpredictable disturbance, such as AGI, rather than to optimize the future by improving prediction. Following Nassim Nicholas Taleb, this direction can be called antifragility, the property of gaining from disturbances rather than merely withstanding them.

[41] Takahashi 2026a, Chapter 6, which rejects the argument of effective accelerationism, the movement that calls for accelerating technological development without restraint, as an inference from what is to what ought to be. The supplementary essay applies the same point to slowness (Takahashi 2026b, Section 11).

An antifragile society learns from its errors while its members remain vulnerable.[42] Section 11.2 states what such a claim must specify and under which conditions plurality supports adaptation.

The claims of this section are empirical and conditional, whereas the protection of plural relationships and contexts rests on the normative premise (Section 2). Given the limits of knowability, no single preparation can be shown in advance to be optimal against all unforeseen shocks (Section 2). A single objective function can still include exploration and redundancy among its policies (Section 5.4). The danger grows where an objective values future options narrowly, resources for exploration are cut, the authority to correct is concentrated, and implementations share dependencies. Alternative routes for trying a different response can then disappear. Shared dependencies then become a single point of failure for every domain that relies on them.

The strategy of protecting adaptive capacity responds to longtermism, here the view that reducing existential risk takes priority over all other moral tasks because of the welfare of the very many people who may exist in the future. In this characterization, existential risk is the risk of human extinction or of an irreversible loss of humanity's future potential.[43] The characterization describes a strong form of the view, and the critique below is narrower still, targeting a strategy with two features. Under deep uncertainty, where estimates of probability and value are unstable, the strategy ranks interventions by expected value without testing whether the ranking survives changes of model. It then turns that ranking into an irrevocable concentration of authority, or into a ground for trading the core of standing of present and future people for computed benefits. Expected-value reasoning as such and longtermism in general are not the target. This paper locates the wrong in trading the core of standing (Section 4), not in the weight given to future people.[44] Non-domination supplies a separate objection to an irrevocable concentration of authority (Section 5.2). A further difficulty is epistemic. In the form adopted here, the more finely one ranks specific interventions, the more one depends on predicting their consequences, which bounded rationality makes hard. Staking the future of civilization on a single ethical calculation, agent, and institutional design is most fragile under bounded rationality (Takahashi 2026a, Chapter 5).

The longtermist philosopher William MacAskill defines value lock-in as an event that causes a single value system, or a set of such systems, to persist for an extremely long time. He contrasts narrow-minded, parochial, and unreflective values with open-minded, ecumenical, and morally exploratory ones (MacAskill 2022, Chapter 4). A longtermism that favors the latter can value exploration and the preservation of future options. Constitutive pluralism shares its concern with existential risk and future generations but differs in its normative constraints. What must not be decided includes binding future participants by a final determination of value (Section 2), and the threshold holds however large the ordinary benefits (Section 4).

Physical ceilings on computation and communication make a permanently unified agent hard to maintain, but they do not reduce the risk of AGI (Takahashi 2026a, Chapters 4 and 6). Capability and influence can concentrate during the transition in ways that become hard to reverse, and this paper

[42] Takahashi 2026a, Chapters 1 and 11. The book treats AGI as the greatest such disturbance. On antifragility, see Taleb (2012).

[43] Takahashi 2026a, Chapter 5, which discusses Bostrom (2014) and Ord (2020). For assessments from within the debate, see Tarsney (2023) and Thorstad (2026). The text narrows the book's second difficulty, the need to predict branch points of civilization, to the ranking of specific interventions, since general provisions such as redundancy and correctability need no such predictions.

[44] The first difficulty that the book raises for longtermism is that weighing the undetermined future of people not yet born against the concrete interests of the present generation is not self-evidently legitimate (Takahashi 2026a, Chapter 5).

treats that concentration as an additional risk. Physical distribution, functional integration, and political concentration are separate variables. A single agent can hold the resources, update authority, and final goals of physically distributed computation (Takahashi 2026c). Catastrophe scenarios can be sorted on two axes, whether further effort may avert them and how probable they are. Institutional design concentrates on scenarios that further effort may avert and whose probability is not negligible.[45] This sorting helps to find where further effort may change outcomes, such as the advantage of attack, which needs one breach while defense must cover every front. Ranking interventions among these scenarios still requires comparing harms, costs, effects, and uncertainty.

8 Premises and what remains if they are rejected

The arguments of Parts I and II rest on normative premises that the paper states rather than derives, and what remains when each is rejected differs. The normative premise rests on equal consideration and the justification owed to those affected, and its reasons do not bind those who reject both commitments (Section 2). Beyond the argument from justification, the threshold for the core of standing rests on a premise of partial aggregation and on a judgment of the core's strength (Section 4). An account that rejects them can permit the loss of the core to be compensated by sufficiently large ordinary benefits, and the disagreement then concerns compensability (Section 4). Non-aggregation rests on a reading of equal consideration in its remaining case, and a further premise gives the prohibition the threshold's strength in that case and in the second kind of case (Section 5.3). A reader who rejects the reading can still accept the prohibition in the two kinds of case that the normative premise rules out directly, where review on the domain's reasons is lost or where the use covers the domains of life. A reader who rejects the further premise can still treat the prohibition in the second kind of case and the remaining case as a strong reason that ordinary benefits must publicly outweigh. The account of where relational value arises is the paper's own, and the institutional reason for protecting relational value holds without it (Section 3.1.3).

Part II. Institutions

Part II asks what the principle requires of institutional arrangements. Using the core of standing and its threshold (Section 4), the three pillars (Section 5), and autonomy of pace (Section 6), it compares arrangements (Section 9) and states the conditions under which the principle can be realized (Section 10).

9 Comparing institutional arrangements

The cases in Sections 9.1 to 9.6 are analytical constructions that test the distinctions and limits of the principle, Section 9.7 turns to catastrophic risk, and Section 9.8 is an illustration. They do not report observed effects or establish that the proposed responses will work.

[45] The book sorts catastrophe scenarios into the unavoidable, those that further effort may avert, and the unlikely (Takahashi 2026a, Chapter 1). Because a scenario can be unlikely and avertable at once, the text separates the two axes.

The comparison uses five criteria. The first is welfare, which includes ordinary benefits and protection from serious loss. The second is effective standing, together with the conditions under which values form, including the relational processes of Section 3.1. The others are the capacity to correct decisions, the capacity to reach and implement binding decisions, and the fairness and sustainability of burdens. The second criterion is defended rather than neutral, and a welfare-based account should be assessed in its own terms. Correction concerns whether and how quickly decisions and standards can change, whereas who can demand those changes falls under standing. An effect that bears on several criteria is listed once, with those criteria, to avoid double counting.

Competing arrangements are assumed to have the same technology, total resources, administrative competence, and degree of goodwill. Centralized administration can include independent review and strong rights, and polycentric administration, in which several centers of decision interact, can include shared infrastructure and central minimum standards. Neither arrangement therefore receives credit merely for its organizational form.

The comparison first applies two constraints. Under the threshold constraint of Section 4, an arrangement that removes the core of standing solely to increase ordinary benefits is inadmissible. Under non-aggregation, an arrangement that uses a comprehensive cross-context evaluation of a person to bind decisions on eligibility or treatment is inadmissible, except under the exception conditions for basic interests. The cases identify conclusions that depend on the threshold of Section 4 and those that also depend on its extension in Section 5.3. Among admissible arrangements, if one does better on the second criterion and no worse on the other four, including ordinary welfare, the defended value of standing supports its adoption. Conversely, if a rival does at least as well on the second criterion, worsens none of the others, and improves at least one, the rival should be supported. A reduction in ordinary welfare counts as a tradeoff even below the threshold of serious loss. Where tradeoffs remain, the conclusion may be a conditional or partial advantage rather than an overall ranking.

9.1 Beneficial provision and control over purposes

Return to the public service of the introduction, which supplies accurate recommendations but excludes challenges to the range of available lives. An auditor can verify whether it applied an eligibility rule correctly. Such verification does not answer whether affected people can revise the rule, the set of educational options, or the terms of delegation. The missing pillar is non-domination, because authority over these purposes escapes the effective control of those who depend on the service. The additional function is an effective route to contest those matters. Information, representation, and feasible alternatives may support this route. This function does not require each person to review every routine decision, and justified collective authorization or delegation remains appropriate.

A strong basic-rights or republican alternative can require the same routes, and the institutional judgment is then equivalent. A welfare account that treats self-formation as part of welfare can also reject the service on its own grounds, so the case shows only that output auditing is insufficient. Those routes may reduce resources for beneficial services, and the costs count. If a safeguard would eliminate basic healthcare and no less costly arrangement can supply that safeguard, constitutive pluralism cannot select it merely because it strengthens standing.

A stronger rival keeps independent review but gives affected people no standing to initiate it. An ex officio review body examines interests, effects on minorities, and conflicts of interest on its own initiative. Suppose that this body is accurate and that removing the right to initiate review yields a net

gain in ordinary benefits, even after counting the errors that such a right could expose. Constitutive pluralism still requires the right to initiate, which belongs to the core of standing. Without it, the affected person's reasons enter review only when the reviewer so decides, and the terms of accountability remain within another body's discretion. A welfare account that lets the gain compensate for this loss of standing can prefer the ex officio design. An account that counts the right to initiate as a welfare component with non-compensable priority agrees with constitutive pluralism. The disagreement therefore concerns compensability rather than welfarism as such, and it rests on the argument for the threshold in Section 4.

9.2 Many providers and one comprehensive ranking

Suppose that separate housing, education, and support institutions use the same comprehensive contribution score, computed from a person's employment, care, learning, and civic conduct. Care and learning then affect access to resources only through the score. Objections can correct the data and the weights, but no route returns a claim to the reasons of an institution's own purpose.

The arrangement violates non-aggregation, because it uses a comprehensive cross-context evaluation of a person to bind decisions on eligibility. The violation depends neither on the number of providers nor on procedural defects. A claim to housing is assessed on overall contribution rather than on reasons that bear on the purpose of housing allocation, and the person loses the standing that Section 5.3 describes. The case assumes that this purpose is to meet need. What the purpose should be is a question for that domain's principles of allocation, not for non-aggregation. Naming the rating of the person as the purpose, however, does not escape non-aggregation (Section 5.3). Within that limit, non-aggregation requires that the reasons that decide a claim bear on the purpose. Because every institution uses the same score, a poor record in one workplace also follows the person into every application for housing, courses, and support. The person therefore cannot escape the terms of the score in any of these domains, and where no other viable context remains, cannot choose again elsewhere either (Section 5.3). In the decisions that the score governs, care is neither recognized as care nor compared on terms that the practice of care supplies (Section 3.1).

The prohibition may seem too strong, because context-specific criteria can increase administrative cost, inconsistency, and opportunities for manipulation. Those costs must be recorded. However, the prohibition is narrow. It leaves available data linkage and aggregation for a limited purpose, and it targets only the binding use of a comprehensive evaluation of a person. Administrative savings do not justify that use unless avoiding it would sacrifice basic interests, and the exception conditions of Section 4 then apply.

A shared need index shows where the line falls. Such an index may draw on the same records of employment, care, and learning as the contribution score. It is acceptable under three conditions. Each agency justifies its variables and weights for its own purpose. It can depart from the index for reasons relevant to that purpose. It keeps its refusals to depart open to review. The index then counts as aggregation for a limited purpose, because it measures need rather than a person's worth across contexts, and need is relevant to each allocation. What separates the index from the contribution score is not the data but the object of evaluation and the justification of each input (Section 5.3). The power to depart and the review of refusals are further safeguards that let reasons tied to each purpose correct the result. However, if a score in practice settles ordinary treatment, discretion to depart from it does not by itself remove its binding role. Review across providers should examine whether stated purposes

actually govern allocation, because a common ranking that recurs despite materially different needs can signal a comprehensive evaluation in practice.

9.3 Public benefit from limited information linkage

Suppose that allocating hospital capacity during an epidemic requires combining clinical records with information about regional movement. Refusing all linkage could cause harmful duplication and omissions, so the comparison must include the harms of both using and withholding the information. Linkage for this limited purpose is compatible with non-aggregation if affected people can learn how the results are used and contest that use. The use remains subject to examination of its necessity, access, retention, quality, and role in decisions. An individual request for exclusion does not establish a general veto. Repurposing the data would violate non-aggregation if they were rebuilt into a permanent ranking that rates the person across aspects of life and settles ordinary employment and housing decisions (Section 5.3). Using the same limited risk measure for another purpose for which it is justified would be a different case. Strong privacy protections can impose equivalent conditions, so the case marks the boundary of non-aggregation rather than a deficiency of privacy theory.

9.4 Cumulative updates without a single controller

Suppose that a workplace, a school, an insurer, and a public agency each modify their procedures within the same season. Each change comes with substantive notice of what will change and why. The public agency uses school records to set support levels, but this linkage serves a limited purpose that has been justified and can be contested. Each organization also allows a response period that would be effective if its change were the only one. Yet the same household must respond to all four while providing care. Within the overlapping deadlines, its members cannot obtain advice and prepare four responses. Where the organizations use AI systems that update rules continuously, such overlaps can become more frequent, although the same kind of system can also help a coordinating body detect them. Judged organization by organization, no organization exercises uncontrolled authority over the household, and no single procedure is defective. Under the actual overlap, however, the household cannot use the routes that would make each organization's power controllable, so non-domination read as effective control may find a violation here as well. Non-aggregation finds none. The loss lies in the timing of the changes relative to the household's capacity to respond, which concerns autonomy of pace, and in the household's lack of usable time, which concerns slack. A third matter is the power to set each program's purposes and to decide which changes to make. That power remains with each organization and is subject to collective control under non-domination, whereas decisions on the timing and sequence of changes and on interim relief fall to a coordinating body under an authorized mandate (see below). Its relief is limited to measures such as extended deadlines and the continuation of previous terms, and the organizations can contest its decisions before an independent review body. What autonomy of pace adds is this diagnosis and a remedy that restores effective control by coordinating timing, without constraining any organization's authority over its own purposes (Section 6).

Three designs show what autonomy of pace adds. In the first, each organization handles appeals separately, so the household would have to prepare four appeals with separate evidence against overlapping deadlines. It has rights but cannot exercise them, and it loses effective access to review, which is part of the core of standing. Under the threshold of Section 4, the lower cost of this design can-

not justify it. In the second, an advice service increases the household's resources for objection but leaves the deadlines unchanged. In the third, the organizations report planned changes to a coordinating body, which detects overlaps for the same households and can grant interim relief, such as extended deadlines or the continuation of previous terms. The contrast between the second and third designs shows that the timing of deadlines matters even when the total time available is the same (Section 6).

Three further designs need no coordinating body to detect overlaps. A general extension lengthens every response period, but delays every change, including beneficial ones, for people who face no overlap. A bundled appeal lets a household contest several changes through one filing with shared evidence. It removes duplicated preparation, but the deadlines still coincide. Provisional continuation keeps a household's previous status once it files a short protective notice, with a later deadline for reasons and evidence. It eases the pressure of coinciding deadlines, but delays changes for every filer and can invite strategic filing.

The third design's distinctive cost lies in detecting overlaps. The organizations must report planned changes and deadlines. The body must then either match affected households across organizations, which requires limited data linkage, or rely on households to report overlaps themselves, which returns part of the burden to them. The third design therefore does worse than its rivals under four conditions. Overlaps may be too rare to repay detection. They may instead be common enough that a general extension costs little more. The burden may lie mainly in duplicated preparation, which a bundled appeal removes. Rescheduling may also cost others more than it saves the household. Measuring cumulative burden can itself become surveillance. Measurement should therefore be limited to relevant tasks, deadlines, and response costs, and it must not grow into a comprehensive evaluation of the household that decides support or relief for its members (Section 5.3).

Duties that no single organization can discharge also need addressees. They are allocated according to each party's contribution to the cumulative change, the range that it controls, and its authority and resources. A merely affected household is not made responsible for its own delays. The coordinating body's authority should extend to the timing and sequence of changes and to interim relief, not to the purposes of each program. Because temporal remedies redistribute benefits among users, the coordinating body's mandate itself requires authorization by an accountable body (Section 3.2). The coordinating body itself must also be contestable.

Where costs passed on through prices, premiums, and taxes ultimately impair someone's effective exercise of standing, the allocation constraint of Section 4 applies. Voluntary coordination can also erode, because organizations that stagger changes or fund coordination may lose ground to competitors that do not. Cross-sector requirements with pooled funding reduce this pressure, although competitors outside those rules remain. The third design, alone or combined with bundled appeals and provisional continuation, therefore has at most a partial advantage. The advantage holds where overlapping deadlines cause the loss and none of the four conditions applies. The design's costs and delays should be stated as tradeoffs.

9.5 Continuation, change, and exit

Consider a small coastal community whose members maintain a school and a practice of building boats together. Older members teach the craft to children through shared work, and each generation changes it. In the sense of Section 3.1, the value of the practice is borne by this continuing interaction,

and no member can sustain it alone. A regional AI planning system compares schools by travel time and measured learning outcomes. It recommends merging the school into a larger one in a nearby town, which offers more subjects and a wider range of peers. Some members wish to continue, others seek the change, and some wish to leave. Children, people who depend on care, and affected outsiders also have interests at stake, and a representative's preference for continuity cannot be treated as every member's.

If the school closes, the practice may lapse. No one would prohibit it, but the occasions of shared work could disappear as each household makes a reasonable choice. What would be lost is not a technique that a manual could record but the shared work in which older members and children respond to each other. Reported satisfaction need not fall, no one's rights need be violated, and the practice that the next generation would have formed has no organized claimant. The planning system does not record this loss, and the comparison must. An evaluation that tracks only households' reported welfare, and treats it as fully compensable by money, would find nothing further to weigh once each household had been compensated. Counting the good that the practice bears leaves a pro tanto reason for public judgment even after such compensation, and even if the larger school would secure the children's opportunities to reconsider (Section 3.1). Welfare theories that count relational achievements, or an evaluation of the practice as a public good, may reach the same reason (Section 3.1.1). The foundation's work is to keep the comparison from stopping at payments to households. The reason confers no claim to continuation, subsidy, or participation. Its weight is assessed against the costs of keeping the school, the freedom of members to leave, and the children's interest in a wider curriculum and peer group. Where continuation obstructs the children's effective access to relationships and perspectives outside the community, it has relational costs as well. Blocking exit is a separate harm to those who wish to leave.

The principle does not decide between continuation and change. Members who prefer the familiar relationships of the community are not failing to exercise autonomy, and members who leave owe no one their continued participation. Once the children's access to other relationships and perspectives is sufficient, the principle gives no further reason to favor the merger merely because encounters in the town might be transformative. What the principle requires concerns the decision and its pace. Independent representation of children, transition assistance, and review of effects on outsiders can support a time-bounded collective decision. The decision's timing must permit effective involvement. Support for carrying the practice into new settings should be considered where feasible and justified, since no payment to individual households replaces the shared work. Those on the losing side should retain access to reasons, avenues for challenge, remedies, and opportunities for future participation. Indefinite deliberation is not neutral, because autonomy of pace treats imposed stasis and imposed change alike as harms. Where transition costs placed solely on those who leave would impair their effective standing, financing or assistance must prevent that impairment.

9.6 Emergency action and the cost of waiting

During a flood, waiting for every affected person's prior agreement can increase serious harm. People who request delay and people endangered by that delay both have claims to consideration, so autonomy of pace does not require universal prior consent. A justified emergency mandate can permit immediate evacuation or traffic restrictions if its scope, responsible decision maker, reasons, duration, and conditions of extension are specified. Such a mandate postpones the prior procedures through

which standing is exercised, but it does not remove standing. Objections remain possible. Expedited independent review can modify or stop a measure without automatically suspending it. The declaring authority should not decide alone when review begins. A prolonged crisis therefore requires, after a publicly justified cumulative period, a transition either to ordinary rules under which those affected take part directly or through representation, or to the end of the exception. A rescue that must proceed without these protections leaves a deficit, which should be recorded and remedied. The record does not settle whether the rescue is permissible (Section 4). Whether under a mandate or in such a rescue, compulsion affecting a person's body requires a separate assessment of the specific power to refuse (Section 3.2).

9.7 Catastrophic risk and coordinated action

Some of the scenarios on which institutional design concentrates (Section 7) need more than plurality. Plurality alone cannot stop a single defecting agent where attack dominates, or an agent built outside a network of mutual monitoring, which directly covers only its participants. The balance of attack and defense is, however, a policy variable that the domain, defensive investment, audit, and the allocation of authority to stop can change (Takahashi 2026c; 2026d, CP-25).

Common safety constraints and rapid coordinated action, as in deflecting an asteroid, are compatible with the principle. It permits justified emergency centralization under the conditions of Section 9.6. Restrictions justified by uncertain catastrophic risks must also meet the conditions of Section 4. Options can be generated plurally while their execution is coordinated, provided that the capacity to generate new frameworks survives the emergency.

9.8 An illustration from the production of knowledge

The production of knowledge illustrates why error correction and standing should be connected without being identified. Questions set by a narrow group can receive reliable answers, and a wide range of questions does not guarantee them. Robert Merton formulated scientific norms as institutional norms that organize criticism and control the influence of private interests, and Helen Longino identifies conditions under which criticism can affect a community's work (Merton 1942/1973, pp. 268–270 and 273–277; Longino 1990, Chapter 4, pp. 76–81). Where AI substantially shapes question formation, research assessment, or changes in methods and infrastructure, the three pillars and autonomy of pace yield three requirements. People affected by research should be able to contest its purposes and evaluation criteria through review independent of the implementing institution, without thereby acquiring researchers' duties. Funding should reserve room for minority hypotheses alongside optimized indicators, which applies slack to research (Takahashi 2026a, Chapter 11). Responsibilities, including for the pace of transitions, should rest with the institutions that control access, funding, and schedules, not with individual researchers alone or with an AI that cannot answer, provide remedies, or bear sanctions.

10 Conditions of realization

Normative justification does not establish institutional feasibility, so evaluation must distinguish the principle, the functions that it requires, and the means that perform them. The necessary functions include a route to contest constituency and purpose, access to relevant reasons, the ability to reach

and implement decisions, remedies for losses, usable time and resources, and control over transfers of evaluation. A principle that protects standing cannot assume that markets will supply these functions, or that creating a public body establishes them.

In public decisions, the functions that the pillars require should be secured before AI's use in them expands in scale and speed. Value choices embedded in large and fast operation are hard to correct afterward (Takahashi 2026a, Chapters 11 and 12). The three institutional principles of Section 5.5 are candidate means for these functions.[46]

Because a provider may ignore a review body's order or comply only in form, a correction route also needs powers that do not depend on the provider's cooperation. Such powers rest on legal authority, independent technical capacity, and political support, which depends on affected people retaining effective influence. Where AI systems supply most technical capacity, independence can erode while legal authority remains. The principle can then diagnose a wrong but cannot by itself supply a remedy.

These functions must also be sustained under cost, competition, and dependence.[47] A right to stop an AI system is effective only if basic interests remain protected during the stop. The required function is therefore a verified route of continuation, degradation, or substitution. Separate systems for the same function, degraded modes that people can run by hand, and rehearsed stopping procedures are candidate means. Their evaluation must include shared dependencies and failures in switching between routes. Because voluntary restraint favors those who do not restrain themselves, safety requirements have to become conditions of market participation and public procurement. Competition across jurisdictions remains a problem of international coordination.

AI systems can help reviewers check facts and execute authorized corrections. When final authority is given only in form, without the substantive ability to judge, such authority is a disguise of responsibility.[48] A review therefore does not count as successful when reviewers cannot verify an AI-assisted assessment. The decision should then remain provisional until a set date for renewed review. To verify an assessment here means to check the facts and reasons on which the decision relies, using evidence or tests independent of the system that produced the assessment. It does not require reproducing the system's computation.

An ecosystem of many AI systems also has failure modes of its own (Takahashi 2026a, Chapter 11), each calling for a remedy that matches its mechanism. Common-cause failure, which concerns how errors are correlated (Section 11.2), calls for independent evaluation on different grounds. Collusion, which requires interaction and incentives that reward coordination, calls for audit of communication between agents and for separating monitors from the monitored (Takahashi 2026c). Systems that each comply with a rule can be combined to circumvent it, so review must cover combinations of systems. Emergent collective behavior can escape existing audit and calls for records that follow processes across systems.

[46]The book names value-setting rights, veto rights, and slack as what should be secured first in public decisions (Takahashi 2026a, Chapter 12).

[47]That these functions can be sustained under competition has not been shown, since, as Appendix 2 notes, the book establishes only necessary conditions and design principles (Takahashi 2026c). Appendix 2 proposes financing the candidate means named in the text as one investment in redundancy.

[48]Takahashi 2026a, Chapter 11. A related analysis is Madeleine Clare Elish's account of moral crumple zones, in which "responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system" (Elish 2019, p. 41).

Part III. Testability

Part III sets out how the empirical claims of Parts I and II can be tested and what evidence would require revising the principle.

11 Testing and revision

11.1 Empirical claims and evaluation

The paper makes empirical claims that evidence can confirm or defeat. They concern practices driven by approval or engagement (Section 3.1.3), whether a mandate to coordinate cumulative burdens can work at an acceptable cost (Section 3.2), the intensifying pressure to measure and optimize relationships (Section 5.1), and the effect of market competition on plurality (Section 5.2). Others concern whether new updates arrive before those affected can complete their responses (Section 6), when the third design does worse than its rivals (Section 9.4), and when plurality supports adaptation (Section 11.2).

Some claims describe conditions of application, such as updates that exceed the capacity to respond and the shaping of options. Indicators of these conditions include viable alternatives, update intervals relative to response periods, and the costs of exit and objection. The paper sets no numerical thresholds and reduces institutional demands where effects are weak. Claims that plural arrangements improve adaptation or the collective correction of errors depend on the conditions stated in Section 11.2, and other claimed benefits must be tested against the mechanisms said to produce them. Evaluation should examine access to distinct evidence, conflicts of interest, and the power to change decisions.

Evaluation distinguishes three objects. Institutional requirements, such as access to alternatives, independent review, resources, and deadlines, can be judged in advance. The achievement of functions, such as the cost of making a claim and the implementation of corrections, is observed afterward. The final outcomes of value formation emerge slowly and depend on much besides institutions. The requirements and the criteria for their satisfaction should be fixed before an intervention. Few complaints can indicate satisfaction, but also exclusion or fear of retaliation. Poor outcomes alone do not show that the requirements were unmet. Criteria may be revised through contestation, provided that the initial criteria and results are retained. No failure can then be redescribed as the absence of an unspecified requirement.

The hypothesis of Section 3.1 about practices driven by incentives for approval or engagement needs the same care. A study should distinguish isolation that increases AI use from AI use that increases isolation, and allow that AI may help people leave harmful relationships or form new ones. It should measure available options, exit costs, and actual participation in relationships separately from satisfaction and time of use, considering the prior state, the alternatives, and the passage of time.

11.2 Conditions under which plurality supports adaptation

This paper treats a claim of antifragility (Section 7) as a hypothesis that must specify the function that gains, the range of disturbances, the baseline of comparison, the time scale, and who bears the

damage. Plurality can preserve a reservoir of frameworks and the capacity to generate new ones under four conditions. The first concerns how errors are correlated. Systems with different designs and training histories can have different blind spots, so that one detects deviations that another misses. However, contributors whose judgments are independent in the sense of Section 3.1 can still fail together when they share data, architectures, assumptions, infrastructure, or update cycles. Evaluation should therefore ask how errors are correlated across tasks and environments, which common causes can make several systems fail at once, and how failures propagate.

The second condition is coordination. In a model by the computer scientists Jon Kleinberg and Manish Raghavan, two firms each hire one candidate from a shared pool, ranking candidates either independently or with a common algorithm. Even where the more accurate algorithm is the better choice for each firm, social welfare, measured by the total quality of those hired, can be higher if both keep their own less accurate rankings.[49] The result concerns social welfare under the model's assumptions, not a general claim that shared algorithms lower individual predictive accuracy, and the model does not show what coordination is needed for adaptation at larger scales. Such adaptation needs institutions that compare several trials, adjust among them, and carry out effective responses jointly, and their effectiveness must itself be tested.

The third condition is institutions that connect perspectives and pass on what is learned, for which slack reserves time and resources (Section 5.4). The fourth condition is containment, meaning that local failures do not propagate into a global catastrophe. Where one of many agents can cause an irreversible loss to all, dispersal and experimentation add danger rather than adaptive capacity. Antifragility is therefore a direction of design only for disturbances that can be contained.

11.3 Failure modes and evidence for reconsideration

The canon, a numbered list of the book's propositions, records conditions for reconsidering the choice of this principle. One condition is a showing that the failure modes specific to plurality cannot be mitigated by institutional design. They are stagnation through the abuse of vetoes, the disappearance of responsibility in distributed structures, entrenched interests disguised as plurality, and delayed response to crises (Takahashi 2026d, CP-6). Stagnation through vetoes is imposed stasis, which autonomy of pace counts as a loss on the same footing as imposed change (Section 6). This also answers the criticism that the principle is a conservatism that freezes existing communities, cultures, and languages (Takahashi 2026c), since the principle protects autonomy of pace rather than present boundaries (Section 9.5). Powers to halt public decisions are therefore domain-specific, and each halting order carries a time limit and re-examination, except for refusals of bodily intervention and similar specific protections (Section 3.2). Repeated objections on one issue can be consolidated into a reasoned final decision, while routes stay open for new evidence and for claims of exclusion from standing. Neither the number of objections nor the objectors' minority status makes them abusive. Against the disappearance of responsibility, the principle pluralizes authority but assigns duties to named parties (Section 9.4). Against entrenchment, non-domination targets existing domination as well as new domination, through periodic audit, sunset clauses, and independent evaluation. Against delay, the principle secures the capacity to carry out choices and permits justified emergency mandates (Sections 6 and 9.6). These responses are design principles, not an operational record. Weighing them

[49] Kleinberg and Raghavan 2021. The result holds for the families of ranking models that meet conditions stated in the paper.

against the failures of unification remains a task of verification.

Another condition, which this paper reads more narrowly than the canon does, is a showing that governance under a single objective function is feasible without loss of value, single points of failure, or uncorrectability. Such a finding need not overcome the limits of knowability of Section 2. A comparison would fix its scope, period, and criteria of loss in advance. Evidence that the expected harms do not arise within that scope would withdraw an empirical reason for preferring plural arrangements there. The normative premise of Section 2 would remain.

11.4 Counterexamples and the scope of revision

Two kinds of counterexample would require revision. In the first, requirements fixed in advance are met, but the protective function that they promise does not arise. Suppose that access to alternative presentation environments (other orderings and displays of options), independent information, resources for inquiry, and review all meet criteria set before an intervention. Suppose also that the ordering and defaults of everyday AI recommendations persistently narrow the alternatives that people imagine. Convergence of choices alone would not establish a loss. The counterexample would require evidence that alternative environments let people discover options that they themselves regard as important, while the original environment removes those opportunities before objection can arise.

No presentation environment is a neutral baseline, because every one mediates what people perceive and consider.[50] The test accepts such mediation and is therefore comparative. It asks whether people retain opportunities to reconsider their options and practical access to alternative environments, and whether their judgments persist after encountering those alternatives. A stable judgment reached after such comparison is evidence of an excluded opportunity if it favors options that the original environment kept out of view. However, whether the closure is wrongful also depends on manipulation in the alternative environment, competing interests, and the feasibility of preserving access. The alternatives should be specified in advance, including some proposed by affected people and independent reviewers. Because the loss can prevent objections from arising, the criterion also asks whether affected people have effective standing to contest the designs that shape their options before anyone objects.

Evidence for the first kind would show that the chosen means fail to produce the promised function, or that requirements derived from the two foundations (Sections 3 and 5) miss part of what the foundations protect. A requirement added afterward to prevent the loss should be recorded as a revision of the implementation requirements or of the principle. Where added requirements cannot resolve the loss without serious competing harm, the principle should not be adopted unchanged.

In the second kind, satisfying the requirements itself produces serious losses or makes collective action impossible. Suppose that the personnel needed for independent review are indispensable for minimum healthcare. Shared review, lower frequency, and narrower mandates should then be compared under the same resources. If none avoids the conflict, the required extent of protection must be reconsidered. As the distinction between justified exceptions and deficits requires (Section 4), the reconsideration must state which protections are sacrificed. A rival that supplies the same protection

[50] According to the Stanford Encyclopedia of Philosophy, the philosopher of technology Peter-Paul Verbeek holds that technologies mediate how people perceive the world and which actions appear possible and desirable to them. He treats the anticipation of such mediations, and design for desirable ones, as a task of design ethics (Franssen et al. 2026, Sections 3.1.2 and 3.2.1). The entry summarizes Verbeek (2011).

more cheaply supports a change of means, not a revision of the principle.

12 Relation to existing theories and to the book

12.1 Relation to existing theories

The concerns behind the three pillars and autonomy of pace have substantial precedents.

The AI researcher Jan Kulveit and colleagues examine gradual disempowerment, a process through which humans lose influence over institutions. They separate local compliance with human instructions from the preservation of human influence across economic, cultural, and political institutions (Kulveit et al. 2025, Sections 2.4 and 5–6). The present argument shares this separation and supplies its own normative reasons for it.

Rawls's account of the basic liberties addresses the development and exercise of moral powers, including the capacity to form, revise, and pursue a conception of the good (Rawls 1982, Sections IV, VII, and XII). It treats resources and institutional conditions as relevant and gives the political liberties the guarantee of fair value used in Section 4. Constitutive pluralism cannot claim to introduce the protection of value formation into this tradition. Its question is how such demands operate when evidence, categories of eligibility, and the time needed to contest decisions depend on several interacting systems.

The political economist Elinor Ostrom studied polycentric governance, in which several centers of decision interact, and arrangements in which participants help make and modify rules, monitoring is accountable, and dispute resolution is accessible (Ostrom 2009, pp. 3–5 and 32–39; 2010, pp. 643–644, 653, and 663–665). Such governance is compatible with central dispute resolution, and it is not a theorem that decentralization always succeeds. For the AI era, this paper translates Ostrom's polycentric order into the requirement that no single AI stack monopolize the medium of collaboration (Takahashi 2026b, Section 10). On the account of Section 5.2, such a monopoly is objectionable where it removes effective control or the grounds for objection. An AI stack combines a foundation model, training data, and a design philosophy.

In their account of democratic experimentalism, the legal scholars Charles Sabel and William Simon treat goals, performance measures, and decision procedures as revisable through local experimentation, reporting, and review (Sabel and Simon 2011, pp. 79–82 and 89–93). Their work on destabilization rights, which permit challenges to institutions insulated from ordinary correction, also discusses continuing participation and challenges raised by constituencies that emerge during implementation (Sabel and Simon 2004, pp. 1052–1053, 1067–1073, 1077–1080, and 1098–1100). Constitutive pluralism therefore cannot be contrasted with an experimentalism that fixes all purposes and excludes new participants.

The remaining questions concern burdens of revision that accumulate across institutions, people's need for time outside an institution's program of continual revision, and conflicts between continuing and changing an arrangement. Constitutive pluralism answers the first and third through autonomy of pace and the second through slack. Where existing theories answer them through their own commitments, the paper's contribution to these protections narrows to synthesis and application.

The philosopher of law and technology John Danaher argues that reliance on algorithmic systems threatens the legitimacy of public decision-making, because such systems constrain human participation in, and comprehension of, those decisions (Danaher 2016, pp. 246, 254, and 266–267). He

weighs the costs of opacity against instrumental and procedural benefits. Constitutive pluralism shares the concern with participation, but it does not weigh the core of standing against ordinary benefits (Section 4).

The design researchers Kars Alfrink and colleagues argue that AI systems should be contestable by design, responsive to human intervention throughout the system lifecycle (Alfrink et al. 2023, pp. 614–615 and 617). The legal scholars Margot Kaminski and Jennifer Urban argue for an individual right to contest AI decisions, modeled on due process and complementing ex ante systemic governance (Kaminski and Urban 2021, pp. 1957–1958 and 1984). Constitutive pluralism shares with Alfrink and colleagues the view that contestation should reach design. The right defended by Kaminski and Urban can extend across contexts. The cumulative case in Section 9 examines burdens that accumulate across several institutions, none of which is defective when assessed alone, without claiming that either account omits them.

The political theorist Chantal Mouffe's agonistic pluralism holds that confrontation between adversaries, rather than enemies, is a condition of democracy's existence. It also holds that every consensus is a temporary stabilization of power that entails some exclusion (Mouffe 2000, pp. 16–17). Constitutive pluralism permits a reasoned final decision in a particular case without unanimity among those affected, assessing it in part by whether they could be heard and can still obtain warranted revision. Mouffe's concern extends to the procedures through which reasons are admitted, since decisions about whose objections count are themselves exercises of power. Constitutive pluralism treats those decisions as contestable, as the review of boundaries in Section 3.2 requires. This responds to exclusion without resolving Mouffe's challenge, because contestation does not remove that power.

In a roadmap to pluralistic alignment, Taylor Sorensen and colleagues distinguish aims such as representing a range of reasonable answers, steering responses toward specified perspectives, and matching a population's distribution of responses. The roadmap does not purport to settle whom or what an AI should serve (Sorensen et al. 2024, Sections 3–4, 5.1, and 6.1). A model can generate several perspectives while a single institution retains exclusive power to select objectives, appoint representatives, and determine which objections matter. Non-domination concerns that power.

In *Plurality*, E. Glen Weyl, Audrey Tang, and the Plurality Community describe the idea that gives the book its title as "technology for collaboration across social difference" (Weyl et al. 2024, Section 3-0). This places collaboration across difference at the center of the comparison. Constitutive pluralism asks what standing should belong to the participants in such collaboration and to others affected by it, and how authority over its continuation should be constrained. The civic AI framework of Tang and Caroline Green seeks "AI that the people it affects can check, correct and switch off" through six packs of principles.[51] Its refusal to blend complementary checks into one score shares a concern with reductive evaluation (Tang and Green 2026, FAQ Q13). Non-aggregation, however, more specifically prohibits the binding use of a comprehensive cross-context evaluation of a person in decisions on eligibility or treatment. Civic AI's corrective loop asks "who can find out we are wrong, make us say so, and make it cost us while there is still time to change course." The framework's entrenchment of "the loop itself and the rights floor that keeps the loop open" also converges with this paper's concern with correctability (Tang and Green 2026, FAQ Q16).

Civic AI's FAQ affirms the right of communities to fork the tools and to rebuild elsewhere (Tang

[51] The packs are attentiveness, responsibility, competence, responsiveness, solidarity, and symbiosis. The first four follow the four phases of care that Berenice Fisher and Joan Tronto set out, and the last two extend care to organizations and to keeping systems small and time-limited (Tang and Green 2026).

and Green 2026, FAQ Q13). In civic AI, the people served by a Kami, an AI model that serves one place and answers to its people, agree on a date for ending or review and on who decides what follows (Tang and Green 2026, "Set up your own Kami"). Constitutive pluralism explicitly extends the demand for reasons and review to any continuation that binds others or brings about unavoidable meaningful losses, including the continuation of a community itself (Section 6). The two frameworks can therefore be read together. Civic AI articulates institutional and practical safeguards (Tang 2025; Tang and Green 2026). This paper develops an account of standing and of the conditions under which those affected take part in deciding when arrangements change or continue.

12.2 Relation to the book

Sections 2 to 6 start from the principle as stated in the book and revised in the supplementary essay and the online Appendix 2, which record each revision (Takahashi 2026b, 2026c). The revisions include the standing face of the normative premise, the reading of vulnerability as the possibility of meaningful loss, and autonomy of pace. They also include the division of value-setting authority and the veto into general standing and specific powers, and the separation of the limits of knowability from the premise. This paper adds the threshold for the core of standing and the conditions for justified exceptions to it (Section 4). It argues for the threshold from the justification owed to those affected, a premise of partial aggregation, and a stated judgment of the core's strength. It also adds the institutional argument for protecting relational value, and independence as an account of alterity (Section 3.1). In addition, it grounds non-aggregation in equal consideration, applies it to private decisions that a person cannot avoid, and states as an additional premise that it shares the threshold's strength (Section 5.3). It keeps the essay's efficacy of response as a safeguard of participation.

13 Scope and conclusion

This paper has argued for a principle that protects the conditions under which those who can suffer meaningful loss from decisions can question and choose again. The principle's purposes at the scales of persons, society, and civilization are related but distinct (Table 1). The normative ground lies at the scale of persons, and the purposes at larger scales rest on further reasons, which for civilization are empirical and conditional (Sections 7 and 11.2). The three pillars and autonomy of pace translate these purposes into questions that institutions can examine. The cases show where established protections reach the same conclusions and where centralized decisions remain justified. The principle does not oppose the use of AI, including in public decisions, but requires conditions under which those affected can still question that use and choose again.

The argument does not show that every society can afford a complete realization, that plural institutions always perform better, or that the principle prevents disempowerment in practice. Those claims depend on research on shared failures, coordination, and resource constraints, and on the conditions examined in Sections 10 and 11. Because capability and acceleration may outpace institutional design, the principles of authority allocation, correctability, and reversibility should be settled before concrete institutions, which can then be revised as conditions change.

Use of AI in the preparation of this paper

I prepared this paper with substantial use of AI systems, following the Human-over-the-Loop (HOL) model that I developed in the book (Takahashi 2026a, Chapter 11; 2026d, CP-5). In this model, humans hold the legitimacy layer of setting purposes, making judgments, and taking responsibility, and AI systems carry the speed layer of generating, searching, and checking. I held the legitimacy layer. I set the question and the scope, chose the normative commitments, decided among alternative arguments, and accepted or rejected each proposed change. The principle and its parts are my own claims, which I set out in my earlier work (Takahashi 2026a, 2026b, 2026d). Anthropic's Claude Opus 5.5 and Claude Fable 5.1 and OpenAI's gpt-6-astra model carried the speed layer of drafting, checking citations, simulating referees, and translating the paper into Japanese. I did not treat agreement among models as evidence that an argument is correct. I have kept records of the revision process, including my instructions, the models' proposals and checks, and my decisions to accept or reject them. I am responsible for the content of this paper.

References

Acemoglu, D., & Robinson, J. A. (2000). Why did the West extend the franchise? Democracy, inequality, and growth in historical perspective. *The Quarterly Journal of Economics, 115*(4), 1167–1199. https://doi.org/10.1162/003355300555042

Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2023). Contestable AI by design: Towards a framework. *Minds and Machines, 33*, 613–639. Published online 2022. https://doi.org/10.1007/s11023-022-09611-z

Arrow, K. J. (1951). *Social choice and individual values*. Wiley.

Berlin, I. (2002). Two concepts of liberty. In H. Hardy (Ed.), *Liberty* (pp. 166–217). Oxford University Press. https://doi.org/10.1093/019924989X.003.0004 (Original work published 1958).

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law, 25*(3), 273–291. https://doi.org/10.1007/s10506-017-9214-9

Callard, A. (2018). *Aspiration: The agency of becoming*. Oxford University Press. https://doi.org/10.1093/oso/9780190639488.001.0001

Chan, K. M. A., Balvanera, P., Benessaiah, K., Chapman, M., Díaz, S., Gómez-Baggethun, E., Gould, R., Hannahs, N., Jax, K., Klain, S., Luck, G. W., Martín-López, B., Muraca, B., Norton, B., Ott, K., Pascual, U., Satterfield, T., Tadaki, M., Taggart, J., & Turner, N. (2016). Opinion: Why protect nature? Rethinking values and the environment. *Proceedings of the National Academy of Sciences, 113*(6), 1462–1465. https://doi.org/10.1073/pnas.1525002113

Coeckelbergh, M. (2013). *Human being @ risk: Enhancement, technology, and the evaluation of vulnerability transformations*. Springer. https://doi.org/10.1007/978-94-007-6025-7

Cyert, R. M., & March, J. G. (1963). *A behavioral theory of the firm*. Prentice-Hall.

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology, 29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences, 6*(4), 485–507. https://doi.org/10.1007/s11097-007-9076-9

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

Forst, R. (2014). Two pictures of justice (C. Cronin, Trans.). In *Justice, democracy and the right to justification: Rainer Forst in dialogue* (pp. 3–25). Bloomsbury Academic. https://doi.org/10.5040/9781472544735.ch-001

Franssen, M., van de Poel, I., Klenk, M., & Santoni de Sio, F. (2026). Philosophy of technology. In *The Stanford encyclopedia of philosophy* (Substantive revision of July 10, 2026). https://plato.stanford.edu/entries/technology/

Fuchs, T., & De Jaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. *Phenomenology and the Cognitive Sciences, 8*(4), 465–486. https://doi.org/10.1007/s11097-009-9136-4

Goodin, R. E. (2007). Enfranchising all affected interests, and its alternatives. *Philosophy & Public Affairs, 35*(1), 40–68. https://doi.org/10.1111/j.1088-4963.2007.00098.x

Gunji, Y.-P., & Kamiura, M. (2004). Observational heterarchy enhancing active coupling. *Physica D, 198*(1–2), 74–105. https://doi.org/10.1016/j.physd.2004.08.021

Gunji, Y.-P., Ito, K., & Kusunoki, Y. (1997). Formal model of internal measurement: Alternate changing between recursive definition and domain equation. *Physica D, 110*(3–4), 289–312. https://doi.org/10.1016/S0167-2789(97)00126-7

Gunji, Y.-P., Shinohara, S., Haruna, T., & Basios, V. (2017). Inverse Bayesian inference as a key of consciousness featuring a macroscopic quantum logical structure. *BioSystems, 152*, 44–65. https://doi.org/10.1016/j.biosystems.2016.12.003

Gunkel, D. J. (2018). *Robot rights*. MIT Press. https://doi.org/10.7551/mitpress/11444.001.0001

Hancock, H. (with Dickey, L.). (2014). *Possibilities*. Viking.

Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review, 35*(4), 519–530.

Hayek, F. A. (1960). *The constitution of liberty*. University of Chicago Press.

Hutter, M. (2005). *Universal artificial intelligence*. Springer. https://doi.org/10.1007/b138233

Inoue, T. (2018). *AI jidai no shin bēshikku inkamu ron* [A new theory of basic income for the AI era]. Kōbunsha. In Japanese.

Kaminski, M. E., & Urban, J. M. (2021). The right to contest AI. *Columbia Law Review, 121*(7), 1957–2048. https://columbialawreview.org/content/the-right-to-contest-ai/

Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences, 118*(22), Article e2018340118. https://doi.org/10.1073/pnas.2018340118

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., & Duvenaud, D. (2025). *Gradual disempowerment: Systemic existential risks from incremental AI development* (arXiv:2501.16946v2). arXiv. https://doi.org/10.48550/arXiv.2501.16946

Kurki, V. A. J. (2019). *A theory of legal personhood*. Oxford University Press. https://doi.org/10.1093/oso/9780198844037.001.0001

Leike, J., & Hutter, M. (2018). On the computability of Solomonoff induction and AIXI. *Theoretical Computer Science, 716*, 28–49. https://doi.org/10.1016/j.tcs.2017.11.020

Lizzeri, A., & Persico, N. (2004). Why did the elites extend the suffrage? Democracy and the scope of government, with an application to Britain's "Age of Reform". *The Quarterly Journal of Economics, 119*(2), 707–765. https://doi.org/10.1162/0033553041382175

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press. https://doi.org/10.1515/9780691209753

MacAskill, W. (2022). *What we owe the future*. Basic Books.

Mackenzie, C., Rogers, W., & Dodds, S. (Eds.). (2014). *Vulnerability: New essays in ethics and feminist philosophy*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199316649.001.0001

Matsuno, K. (1989). *Protobiology: Physical basis of biology*. CRC Press.

Matsuno, K. (1996). Internalist stance and the physics of information. *BioSystems, 38*(2–3), 111–118. https://doi.org/10.1016/0303-2647(95)01580-9

Matsuno, K., Gunji, Y.-P., & Rössler, O. E. (1997). *Naibu kansoku* [Internal measurement]. Seidosha. In Japanese.

Merton, R. K. (1973). The normative structure of science. In N. W. Storer (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267–278). University of Chicago Press. (Original work published 1942).

Mill, J. S. (1879). *Utilitarianism* (7th ed.). Longmans, Green, and Co. https://www.gutenberg.org/cache/epub/11224/pg11224-images.html

Mouffe, C. (2000). *Deliberative democracy or agonistic pluralism* (Political Science Series 72). Institute for Advanced Studies. https://irihs.ihs.ac.at/id/eprint/1312/

Nagel, T. (1970). Death. *Noûs, 4*(1), 73–80. https://doi.org/10.2307/2214297

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*, 119–158. https://nissenbaum.tech.cornell.edu/papers/H.%20Nissenbaum,%20_Privacy%20as%20Contextual%20Integrity.pdf

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus, 140*(4), 32–48. https://doi.org/10.1162/DAED_a_00113

Ohya, T. (2021). Society 5.0 to jinkaku naki tōchi [Society 5.0 and governance without personhood]. *Jōhō Tsūshin Seisaku Kenkyū [Journal of Information and Communications Policy], 5*(1), 1–14. https://doi.org/10.24798/jicp.5.1_1. In Japanese.

Ohya, T. (2024). AGI to shakai to jinrui to [AGI, society, and humankind]. *Jōhō Tsūshin Seisaku Kenkyū [Journal of Information and Communications Policy], 8*(1), 19–31. https://doi.org/10.24798/jicp.8.1_19. In Japanese.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Bloomsbury.

Ostrom, E. (2009). *A polycentric approach for coping with climate change* (Policy Research Working Paper 5095). World Bank. https://dlc.dlib.indiana.edu/dlc/bitstreams/ad35eff0-7ba4-48d6-b850-98f08777904f/download

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review, 100*(3), 641–672. https://doi.org/10.1257/aer.100.3.641

Paul, L. A. (2014). *Transformative experience*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198717959.001.0001

Pettit, P. (1997). *Republicanism: A theory of freedom and government*. Oxford University Press.

Pettit, P. (2005). *Two-dimensional democracy, national and international* (IILJ Working Paper 2005/8). Institute for International Law and Justice, New York University School of Law.

https://www.iilj.org/wp-content/uploads/2016/08/Petit-Two-Dimensional-Democracy-National-and-International-2005.pdf

Pettit, P. (2006). The determinacy of republican policy: A reply to McMahon. *Philosophy & Public Affairs, 34*(3), 275–283. https://doi.org/10.1111/j.1088-4963.2006.00068.x

Rawls, J. (1982). The basic liberties and their priority. In S. M. McMurrin (Ed.), *The Tanner lectures on human values* (Vol. 3). University of Utah Press. https://tannerlectures.org/documents/the-tanner-lectures-on-human-values-3/

Raz, J. (2006). The problem of authority: Revisiting the service conception. *Minnesota Law Review, 90*(4), 1003–1044. https://doi.org/10.24926/265535.919

Sabel, C. F., & Simon, W. H. (2004). Destabilization rights: How public law litigation succeeds. *Harvard Law Review, 117*(4), 1015–1101. https://doi.org/10.2307/4093364

Sabel, C. F., & Simon, W. H. (2011). Minimalism and experimentalism in the administrative state. *Georgetown Law Journal, 100*, 53–93. https://charlessabel.com/papers/SabelSimon.pdf

Saijo, R. (2016, November 9). *AI gijutsu ni taisuru ningen no kizutsukiyasusa to jiritsuteki de shinsei na ikikata no kanōsei ni tsuite* [On human vulnerability to AI technology and the possibility of an autonomous and authentic life] [Lecture]. AI Shakairon Kenkyūkai [AI Society Study Group], 18th meeting, Tokyo. In Japanese.

Schiavio, A., & De Jaegher, H. (2017). Participatory sense-making in joint musical practice. In M. Lesaffre, P.-J. Maes, & M. Leman (Eds.), *The Routledge companion to embodied music interaction* (pp. 31–39). Routledge. https://doi.org/10.4324/9781315621364-4

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT Press.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning* (Proceedings of Machine Learning Research, Vol. 235, pp. 46280–46302). PMLR. https://proceedings.mlr.press/v235/sorensen24a.html

Takahashi, K. (2026a). *AGI—人間を超える知能は文明をいかに変容させるか* [AGI: How Superhuman Intelligence Reshapes Civilization]. Kodansha. In Japanese.

Takahashi, K. (2026b). *Supplementary essay 1: Deriving constitutive pluralism*. Published on the author's website in Japanese and English. https://koichi-takahashi.me/en/agibook/ronko-pluralism/

Takahashi, K. (2026c). *Appendix 2: Responses to anticipated criticisms*. Online appendix to *AGI*. https://koichi-takahashi.me/en/agibook/appendix-2/

Takahashi, K. (2026d). *AGI machine-readable companion bundle (OKF): Canonical propositions* (canon v0.6.3). https://koichi-takahashi.me/agibook/okf/propositions/index.en.md

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder*. Random House.

Tang, A. (2025, September 1). *6-Pack of Care: A manifesto*. Civic AI. CC0. https://civic.ai/manifesto/

Tang, A., & Green, C. (2026). *Civic AI: 6-Pack of Care* [Website, amended October 11, 2026]. CC0. Retrieved November 3, 2026. https://civic.ai/

Tarsney, C. (2023). The epistemic challenge to longtermism. *Synthese, 201*(6), Article 195. https://doi.org/10.1007/s11229-023-04153-y

Taylor, C. (1995). Irreducibly social goods. In *Philosophical arguments* (pp. 127–145). Harvard University Press.

Thorstad, D. (2026). The scope of longtermism. *Australasian Journal of Philosophy, 104*(2), 399–420. Published online 2025. https://doi.org/10.1080/00048402.2025.2521030

Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness, 2021*(2), Article niab034. https://doi.org/10.1093/nc/niab034

Uhlaner, C. J. (1989). "Relational goods" and participation: Incorporating sociability into a theory of rational action. *Public Choice, 62*(3), 253–285. https://doi.org/10.1007/BF02337745

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind*. MIT Press. https://doi.org/10.7551/mitpress/6730.001.0001

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Voorhoeve, A. (2014). How should we aggregate competing claims? *Ethics, 125*(1), 64–87. https://doi.org/10.1086/677022

Walzer, M. (1983). *Spheres of justice: A defense of pluralism and equality*. Basic Books.

Weyl, E. G., Tang, A., & the Plurality Community. (2024). *Plurality: The future of collaborative technology and democracy*. Public domain (CC0). https://plurality.net/

Whelan, F. G. (1983). Prologue: Democratic theory and the boundary problem. In J. R. Pennock & J. W. Chapman (Eds.), *Liberal democracy* (Nomos XXV, pp. 13–47). New York University Press.

Wolpert, D. H. (2008). Physical limits of inference. *Physica D, 237*(9), 1257–1281. https://doi.org/10.1016/j.physd.2008.03.040