Corrigible Goodness Under Constraint

Authoritative Renderings, Misfit, Answerability, and Revision in Good-Oriented Systems

David T. Swanson

March 19, 2026

Abstract

This paper argues that under finite, mediated, and fallible conditions, no system can responsibly pursue the good through fixed targets alone. Real systems do not act on the good in unconstrained contact with reality. They act through authoritative renderings: operative targets, proxies, representations, procedures, thresholds, files, scores, and other structures that acquire standing over cases. Because these renderings are selective and structurally vulnerable to misfit, any system that claims to pursue the good must keep them corrigible, answerable, and revisable relative to reality and the subject-processes they govern. The constitutive danger of finite good-oriented systems is therefore not target use as such, but closure in authoritative renderings. *Target closure* is one especially important form of that broader danger: once the renderings through which systems actually govern harden into self-authorizing practical authorities, systems can optimize misfit rather than the good. Correction is therefore not an optional moral virtue but a structural necessity of good-oriented systems. The paper sharpens this claim by distinguishing declared good, authoritative rendering, operative target, and real good; by showing why governance over subject-processes intensifies vulnerability to misfit; and by arguing that corrigibility is not real unless misfit can become detectable, travel through a live challenge path and usable correction channel, and reach a locus with authority to revise what is actually governing the case within a meaningful burden and time profile. The framework does not provide a final metaphysics of goodness or a complete ranking of goods. Its narrower aim is to identify what must be true of systems that claim to pursue the good under finite conditions. The payoff is a stronger account of why revision, traceability, subject standing, anti-closure, and real rather than merely symbolic correction are constitutive conditions of responsible design and governance.

1 Introduction

1.1. Motivating Problem

Consider a familiar case. A public benefits system is designed to support people in conditions of financial instability while preserving administrative regularity, managing finite public resources, and preventing fraud. Its public justification is straightforward enough: direct support to those who need it under fair and workable institutional constraints. But in practice the system does not act on need in all its human and practical complexity. It acts through forms, deadlines, documentation requirements, eligibility thresholds, automated flags, compliance states, and case files. A claimant who is genuinely in need may still be denied because the system successfully tracks documentation rather than need, procedural compliance rather than practical vulnerability, or administrative legibility rather than what is actually goods-relevant for the life it governs.[10, 7, 2]

That kind of problem is not confined to welfare administration. Hospitals claim to promote health. Schools claim to support learning and development. Risk-assessment systems claim to reduce harm. Public agencies claim to serve citizens. AI systems are increasingly justified in terms of safety, benefit, fairness, or welfare. Across domains, institutions and systems present themselves as oriented toward some good, or at least toward some outcome they take to be good enough to justify action.

Yet systems can be sincerely or confidently good-oriented and still go badly wrong. They can optimize the wrong target, preserve the wrong proxy, encode too-thin a model of the subject, or harden a once-useful rendering into an authority it no longer deserves. They can produce orderly outcomes that remain deeply misaligned with the lives they govern. They can declare success in terms the system itself defined while the governed subjects experience degradation, unreadability, humiliation, burden, foreclosed possibility, or systematically distorted treatment.[1, 6]

This problem is often described in one of two inadequate ways. Sometimes it is treated as a problem of bad intentions: the system failed because the actors were corrupt, cruel, or indifferent. Sometimes it is treated as a problem of implementation: the goal was sound, but execution was poor. Both diagnoses can be true in particular cases. But they miss a deeper structural problem. Even systems with decent aims, trained personnel, and apparently reasonable procedures can go wrong because the system pursues the good through renderings that acquire standing over cases and can drift away from what is genuinely good for the beings they govern.

That divergence matters because real systems do not pursue the good by direct grasp. They pursue it through authoritative renderings: targets, proxies, metrics, categories, models, files, thresholds, procedures, and operative representations that become action-guiding over cases. The gap between what the system declares to be good, what its authoritative renderings actually govern by, and what is genuinely good is therefore not an occasional accident layered onto otherwise transparent action. It is a standing structural possibility of finite governance itself. The distinctive danger of finite good-oriented systems is not only bad target choice at the outset. It is the hardening of authoritative renderings into practical authorities no longer open to goods-relevant revision. *Target closure* is one especially important form of that broader danger.

1.2. Central Question

The central question of this paper is simple to state but difficult to answer:

> *What must be true of a system if it is to pursue the good responsibly under finite, mediated, and fallible conditions?*

This question should be read carefully. The paper is not asking only what good systems should aim at. It is asking what sort of structure a system must have if its claim to pursue the good is to remain credible once one acknowledges that the system acts through authoritative renderings over cases rather than through direct, exhaustive contact with what is good. The issue is not only whether a system has the right aim. It is whether the renderings through which it actually governs remain answerable to the good they claim to serve.

1.3. Main Proposal

The paper's main proposal is this:

> *Because real systems act through selective, fallible, scope-bounded authoritative renderings over cases and subject-processes they do not fully exhaust, any system that claims to pursue the good must keep those renderings corrigible, answerable, and revisable. Correction is therefore not an optional moral virtue but a structural necessity of good-oriented systems under finite conditions.*

More sharply, the paper argues that the constitutive danger of finite good-oriented systems is not target use as such, but closure in authoritative renderings. Operative targets are one especially important subtype of such rendering, but not the only one. Files, classifications, thresholds, scores, and other operative structures can also acquire standing strong enough to govern outcomes. The question is therefore not only whether systems choose good targets. It is whether the renderings through which they actually act remain open to goods-relevant revision. Only *real* rather than merely symbolic corrigibility keeps good pursuit from hardening into optimized misfit.

This proposal is deliberately narrower than a full theory of goodness. The paper does not provide a final metaphysics of the good. It does not assume that all goods can be ranked on one scale, nor that every domain permits the same model of participation or revision. Its claim is instead meta-structural. It requires only that declared goods and authoritative renderings are not self-authenticating and may be answerably criticized by effects on governed cases, burden patterns, subject reports, rival interpretations, and other correction-worthy forms of misfit. Whatever the good ultimately is, a system cannot responsibly claim to pursue it if the renderings through which it actually governs are closed against revision in light of such misfit.

1.4. Distinctive Contribution

The paper's contribution should also be stated in relation to adjacent work. *Mediated Judgment Under Constraint* analyzes how systems govern through operative representations that acquire standing and decisional force over cases. *Corrective Ethics Under Constraint* analyzes why mediated authority conditions are ethically assessable and why answerability, burden sensitivity, and correction responsibility are ethically load-bearing.[16, 14] A related bridge object on rendering and authority explains how renderings come to acquire standing over judgment and action. The present paper addresses a different downstream object: what must be true if systems that claim to pursue the good are to govern through such renderings responsibly at all.

Its distinct contribution is therefore not merely to say that systems should be accountable or revisable. It is to identify closure in authoritative renderings as a constitutive danger of finite good-oriented systems and to show that *real corrigibility* at the level of those renderings is the structural condition that keeps good pursuit from hardening into optimized misfit. The paper's strongest claim is not simply that revision is desirable. It is that systems which claim to pursue the good through authoritative renderings are not responsibly intelligible unless those renderings remain answerable and really revisable rather than merely reviewable in form.

1.5. Roadmap

The paper proceeds as follows. Section 2 situates the proposal against nearby but insufficient approaches. Section 3 defines the paper's key terms and distinctions, including declared good, authoritative rendering, operative target, real good, misfit, and real corrigibility. Section 4 states the main theory of corrigible goodness under constraint, including the middle mechanics of real corrigibility: misfit detection, challenge paths, correction channels, revision authority, burden, and latency. Section 5 explains why this theory is needed and what confusion it resolves. Section 6 shows explanatory payoff through institutional and technical cases, with a fully worked benefits-administration example. Section 7 addresses major objections. Section 8 clarifies scope, limits, and visible residue. Section 9 sketches implications and future work. Section 10 concludes. Appendix A offers provisional definitions, and Appendix B provides a compact formal sketch.

2 Background and Rival Views

This section does not reject the neighboring approaches just discussed. Each captures something important. The question is what each leaves underdescribed for the purposes of this paper. The argument of *Corrigible Goodness Under Constraint* is not that value theory, optimization theory, or corrective and participatory approaches are mistaken. It is that none by itself fully isolates the structural problem of good-oriented systems pursuing the good through authoritative renderings under finite conditions.

2.1. Substantive Good-First Approaches

One familiar family of views begins by identifying some substantive account of the good and then treating action-guidance as downstream from that account. On this picture, the main philosophical burden is to say what is genuinely valuable, what flourishing consists in, what goods matter most, or what ultimate end should orient practical reason. That work is indispensable.[12, 8] Without some account of value, design easily collapses into empty optimization.

But a substantive theory of the good does not by itself tell us what a live system must be like if it is to pursue the good responsibly under finite conditions. Even a strong value theory leaves open the fact that real systems act through mediation, proxy selection, partial information, institutional lock-in, and design tradeoffs. It may tell us what matters without yet telling us how a finite system should remain answerable when the authoritative renderings through which it actually governs drift away from what is genuinely goods-relevant.

The point, then, is not that substantive value theory is dispensable. It is that value theory alone does not solve the downstream structural problem addressed here: how systems that claim to pursue the good can do so without allowing the renderings through which they act to harden into self-authorizing authority.

2.2. Outcome-First and Optimization Approaches

Another family of views starts from outcomes and asks what systems should optimize, maximize, or secure. This approach has obvious practical force. Systems need targets. Policy requires metrics. Organizations must encode procedures. Public institutions often cannot function without some declared expected outcome around which coordination is organized.

What this picture gets right is that systems require design targets. What it misses is that targets are only one species of a broader problem. Systems do not govern through targets alone. They govern through authoritative renderings: files, categories, scores, thresholds, classifications, and other operative structures that acquire standing over cases. A declared outcome is therefore not identical with the good simply because the system was built around it. A system can optimize an encoded target while moving further away from what is genuinely good for the governed case or subject-process, and it can also allow other renderings besides targets to acquire more authority than their warrant can bear. It can become highly competent at pursuing a distorted rendering of the good.[1, 6, 13]

The problem is therefore not only which outcome should be selected. It is also whether the authoritative renderings through which the system acts remain revisable in light of misfit, whether they continue to answer to what matters, and whether rendering success is being mistaken for successful pursuit of the good. This is where the paper locates its central danger. The distinctive risk of finite good-oriented systems is not merely bad outcome choice at the outset. It is closure in authoritative renderings, of which *target closure* is one especially important case.

2.3. Corrective and Participatory Approaches

A third family of views emphasizes voice, participation, appeal, audit, democratic revision, or other forms of correction. This family gets something crucial right: systems that govern persons should not be wholly sealed off from the people they affect, and revision matters wherever power is exercised through abstraction.

What these approaches often leave underdeveloped, however, is the deeper structural reason correction matters. Correction is not merely a good political norm or a morally attractive institutional feature. It is necessary because good pursuit under finite mediation is structurally vulnerable to misfit at the level of the authoritative renderings through which the system actually governs. In other words, correction should not be defended only as fairness, participation, or accountability in a broad sense. It should also be defended as a constitutive requirement of any responsible good-oriented system.[3, 4]

This point matters especially because systems can simulate openness while remaining closed in practice. They can provide review, appeal, or complaint in form while denying real revision authority, imposing prohibitive burdens, or allowing only case-level exceptions while leaving the governing rendering intact. For that reason, the question is not only whether revision exists, but whether corrigibility is real rather than symbolic, and whether it reaches what the system is actually using to govern.

2.4. The Paper's Position

The present paper therefore occupies a middle position. It does not replace substantive value theory. It does not deny that systems need targets. It does not claim that participation solves everything. Its central claim is narrower and more structural:

> *Once one recognizes that good pursuit is always mediated, finite, and vulnerable to misfit, correction becomes a constitutive condition of any system that purports to aim at the good.*

More sharply, the paper argues that the constitutive danger of finite good-oriented systems is closure in authoritative renderings: once the renderings through which systems actually govern become self-authorizing, systems can optimize misfit. *Target closure* is one especially important form of that broader danger. Only real rather than symbolic corrigibility keeps good pursuit from hardening into distortion, domination, or delusion.

The paper therefore does not begin by solving the entire question of what goodness is. It begins by asking what must be true of systems that claim to pursue the good without allowing the authoritative renderings through which they govern to harden into unquestionable authority.

3 Core Definitions and Distinctions

This section fixes the main terms of the argument. The aim is not to multiply vocabulary for its own sake, but to keep several recurrent confusions from collapsing into one another. In particular, the paper depends on distinguishing what a system says it is pursuing, the renderings through which it actually governs, what it optimizes within those renderings, and what is genuinely good for the cases it governs.

3.1. Good-Oriented System

A *good-oriented system* is any system that claims, explicitly or effectively, to pursue, protect, optimize, preserve, or realize some good. The category is intentionally broad. It includes institutions, bureaucracies, technical systems, governance systems, and hybrid socio-technical arrangements. The point of the term is to identify systems whose practical authority is justified, at least in part, by the claim that they are moving toward something good rather than merely producing some outcome whatsoever.

This term is needed because the paper is not about all systems in general. It is about systems whose claims to authority depend on some relation to the good.

3.2. Declared Good, Authoritative Rendering, Operative Target, and Real Good

A *declared good* is the good as stated, avowed, justified, or publicly named by the system. It is what the system says it is aiming at.

An *authoritative rendering* is the operative structure through which a system actually governs a case with enough standing to shape outcomes. It may take the form of a file, category, score, threshold, model, workflow state, proxy architecture, or other rendering that becomes action-guiding over the case. The term matters because systems do not govern through declared value alone. They govern through renderings that acquire decisional force.

An *operative target* is one especially important subtype of authoritative rendering: the encoded objective, proxy, metric, criterion, threshold, or rule through which the system organizes pursuit of its declared good. This distinction matters because targets are not the whole middle object of the paper. Systems govern through a wider class of authoritative renderings, of which targets are one particularly important form.

A *real good*, by contrast, is what is genuinely goods-relevant for the governed case or subject-process. This paper does not provide a final metaphysics or complete ranking theory of the real good. Its claim is narrower but still robust: declared goods and authoritative renderings are not self-authenticating, and they can be answerably criticized by effects on governed cases, burden patterns, subject reports, rival interpretations, and other correction-worthy forms of misfit.[1, 6]

This structure is one of the paper's most important. It blocks the common slide from "what the system says is good" to "what the system operationalizes" to "what is actually good." It also

prevents operative targets from being mistaken for the entire governing layer.

3.3. Subject-Process

A *subject-process* is a finite, self-maintaining, world-coupled process-pattern for which lived significance may be relevant to adequate governance. The term is meant to block a common flattening move. Systems often treat governed beings as administrable cases, profiles, risk-bearers, students, patients, clients, or user-types. Those renderings may be operationally necessary. But where the governed object is a subject-process, the system's thin rendering may fail to preserve what matters about the case as lived.[15, 3]

This term should be used in a bounded way. It is not the sole bridge of the paper's argument. The core theory applies to good-oriented systems generally. *Subject-process* matters because it intensifies the adequacy burden in domains where lived significance may be relevant to what is genuinely good.

3.4. Warranted Scope and Operative Scope

The *warranted scope* of an authoritative rendering is the bounded domain within which it is justified, informative enough, or reliable enough to guide action responsibly. The *operative scope* of a rendering is the domain across which the system actually uses it to classify, route, permit, deny, prioritize, or intervene.

This distinction matters because the danger in the paper is not only that renderings are selective. All renderings are selective. The deeper danger is that a rendering's operative scope can outrun what its warrant can bear. A target, score, file, or category may begin as a bounded aid to action and later become authoritative across a wider range of cases or decisions than it can adequately carry.

3.5. Misfit and Misfit Signals

Misfit is the divergence between declared or rendered good pursuit and what is genuinely goods-relevant for the governed case or subject-process. Misfit should be distinguished from mere deviation. A subject can diverge from a target because the target is badly designed. Not every divergence therefore counts as system success thwarted. Sometimes it is evidence that the rendering itself is wrong, too thin, or operating beyond its warranted scope.

A *misfit signal* is a complaint, evidence pattern, experiential report, outcome pattern, external check, or failure indicator through which misfit becomes detectable. Corrigibility requires more than the abstract possibility that something could be wrong. It requires that misfit can in fact become hearable, legible, or institutionally visible. Systems may also shape what is allowed to count as a signal, which is one reason closure can become self-reinforcing.

These terms are needed because the paper's argument turns on more than the mere possibility of error. It turns on whether goods-relevant error can actually show up inside a system strongly enough to matter.

3.6. Correction, Corrigibility, and Correction Types

Correction refers to revision, recalibration, override, redesign, suspension, rollback, or other adjustment in response to misfit. *Corrigibility* is the structural capacity of a system to remain revisable relative to such misfit.

Not all correction is of the same kind. A system may permit:

- **case correction:** revision of a particular outcome,
- **target correction:** revision of the encoded objective, metric, proxy, or threshold,
- **rendering correction:** revision of the authoritative rendering through which the case is governed,
- **architectural correction:** revision of the broader system design,
- **suspension or rollback:** halting or reversing a process that is generating systematic goods-misfit.

This distinction matters because a system may permit isolated case exceptions while leaving its governing rendering or target architecture effectively closed. In that situation it appears responsive while remaining structurally unchanged.

3.7. Challenge Paths, Correction Channels, Revision Authority, Burden, and Latency

A *challenge path* is the route by which an authoritative rendering can be questioned, contested, revised, overridden, or displaced. A *correction channel* is the actual path by which a misfit signal can alter a case outcome, target, rendering, or system procedure. A *revision authority* is the actor, office, role, or mechanism with power to enact meaningful correction rather than merely record concern.

Correction burden names the labor, delay, cost, expertise, exposure, or risk required to activate correction. *Correction latency* is the lag between detected misfit and meaningful revision.

These concepts matter because reviewability is not the same thing as corrigibility. A system may allow appeal in form while requiring prohibitive burdens, imposing long delays, or routing challenge to actors with no power to change what is actually governing the case. In such cases, correction exists ceremonially rather than structurally.[14, 7]

3.8. Answerability

Answerability is the condition in which a system remains ethically and practically answerable not only to its own declared aims or internal metrics but to reality and to the subject-processes it governs. A system is not answerable merely because it performs well by its own scorecard. It is answerable when challenge, revision, and comparison between governing rendering and real effect remain possible in a meaningful sense.

This term is needed because the paper is not only about whether systems can be changed. It is also about what they remain answerable to while they act.

3.9. Optimization Logic and Closure

Optimization logic is the system's operative mode of selecting, ranking, or pursuing outcomes relative to its governing renderings. *Closure* occurs when a system hardens its authoritative renderings, targets, categories, or optimization logic against meaningful revision. The problem is not stability as such. Systems need stable renderings to function. The problem is closure: treating the current rendering of the good as beyond goods-relevant correction.

Target closure is one especially important form of this broader problem. It occurs when an operative target is treated as self-authorizing even though it remains selective, fallible, and potentially misfitting.

This distinction blocks one of the easiest misunderstandings of the paper. The argument is not against having targets or renderings. It is against treating them as self-authorizing.

3.10. Subject Standing

Subject standing is the condition under which the governed subject-process has recognized standing in challenge, revision, or design. The point of this term is not that every subject must directly co-author every system. It is that subjects should not be treated as pure objects of optimization without any structured place in revision, challenge, or reassessment of the renderings that govern them.

This term matters because the paper is not only about correcting systems from above. In goods-thick domains, those governed are often among the main sites where misfit becomes visible.

3.11. Load-Bearing Distinctions

Several distinctions organize the rest of the paper.

The first is the distinction between *declared good*, *authoritative rendering*, *operative target*, and *real good*. This blocks the automatic equation of the system's self-description, governing rendering, or proxy with what is genuinely good.

The second is the distinction between *warranted scope* and *operative scope*. This blocks the silent expansion of a rendering's authority beyond the domain for which it was originally credible.

The third is the distinction between *rendering success*, *target success*, and *good success*. This blocks the substitution of internal performance for successful good pursuit.

The fourth is the distinction between *optimization* and *corrigibility*. This blocks the thought that once a system has a target or governing rendering, maximizing it is enough.

The fifth is the distinction between *correction* and *discretionary patch*. This blocks the idea that

revision is a charitable supplement rather than a constitutive structure.

The sixth is the distinction between *reviewability* and *real corrigibility*. This blocks the thought that formal appeal or complaint channels are sufficient even when they cannot actually alter what is governing the case.

The seventh is the distinction between *challenge path*, *correction channel*, and *revision authority*. This blocks the assumption that a complaint path by itself proves revisability.

The eighth is the distinction between *timely correction* and *delayed correction*. This blocks the claim that eventual revision is always enough when delay itself may nullify correction in practice.

The ninth is the distinction between *subject standing* and *passive subjecthood*. This blocks the design error of treating governed beings as measurable objects with no epistemic or normative standing.

These distinctions do most of the paper's structural work. Together, they make it possible to say why correction is necessary without collapsing correction into the whole of goodness.

4 Main Theory

This section develops the paper's central argument step by step. The basic claim is not merely that systems sometimes fail, nor merely that correction is a good institutional feature. It is that systems claiming to pursue the good under finite conditions do so through authoritative renderings that are structurally vulnerable to misfit, and that this vulnerability makes real rather than merely symbolic corrigibility a constitutive condition of responsible good pursuit.

4.1. Good Pursuit Is Always Mediated Through Authoritative Renderings

The first step of the argument is simple: systems do not pursue the good through direct, transparent contact with reality. They pursue it through mediation. They define targets, select proxies, build categories, choose metrics, construct models, assemble files, and act through operative representations that acquire standing over cases. This is not a contingent defect of badly designed systems. It is a standing condition of finite action.[10, 16]

A hospital does not act on health in all its fullness. It acts through diagnoses, measurements, protocols, records, and triage categories. A school does not act on flourishing as such. It acts through attendance records, behavioral categories, performance signals, curricular structures, and institutional goals. A welfare system does not act on need in pure immediacy. It acts through forms, eligibility criteria, compliance states, thresholds, and administrative summaries. An AI-mediated system acts through input features, internal abstractions, labels, rankings, and output classes.

This matters because every such mediation is selective. What the system pursues is never simply the good as such. It is some encoded, modeled, administrable rendering of it. The basic problem of the paper begins there: good-oriented systems pursue the good through renderings that make action possible only by reducing, organizing, and filtering what is allowed to count. Once those renderings

acquire authority over cases, their selectivity becomes practically and normatively consequential.

4.2. Authoritative Renderings Are Structurally Vulnerable to Misfit

Once good pursuit is understood as mediated, a second point follows. Any authoritative rendering through which a system governs is structurally vulnerable to misfit. This is not merely because people sometimes choose bad proxies out of incompetence or malice. It is because operationalization itself is selective. A rendering can fail by selecting the wrong proxy, preserving the wrong distinction, ignoring relevant context, overgeneralizing across cases, freezing what should remain revisable, or flattening what matters in the case it governs.[1, 6]

The question, then, is not whether a rendering can ever be good enough to use. Of course it can. Real systems cannot function without renderings of some kind. The question is whether a rendering can be treated as self-authenticating once it acquires standing. The theory's answer is no. Because authoritative renderings remain selective and residue-bearing, good pursuit through them remains structurally vulnerable to misfit.

This yields one of the paper's main principles: success against an authoritative rendering cannot by itself certify successful pursuit of the good. A system can be effective relative to its own governing rendering while being systematically wrong about what matters. This is the point at which the paper's central danger first comes clearly into view. The problem is not only that renderings can fail. It is that they can fail while still appearing successful in their own terms.

4.3. Operative Targets as a Special Case

Operative targets are one especially important subtype of authoritative rendering. They matter because many systems organize action through explicit objectives, metrics, thresholds, or proxies that define what counts as success. But targets are not the whole middle object of the paper. Systems also govern through files, classifications, scores, workflow states, categories, and other renderings that may acquire decisional standing even when they are not framed as formal targets.

This broader framing matters because the danger at issue is not exhausted by target-setting alone. A system can give excessive authority to a case file, a risk category, a documentation status, or a ranking output even where no explicit target dominates the whole design. The paper therefore retains target language because it remains important, but treats target closure as one especially important form of a broader danger: closure in authoritative renderings.

4.4. Subject-Processes Intensify the Problem

This vulnerability becomes sharper when the governed object is a subject-process. The core argument of the paper does not depend on every governed case being a subject-process, nor on a final metaphysics of subjecthood. The general theory applies to good-oriented systems as such. But where the governed case is a subject-process, the adequacy burden becomes more demanding because lived significance may be relevant to what is genuinely good.[15, 3]

A school may govern through attendance records, compliance categories, and performance signals while missing humiliation, fear, alienation, or distortion of development. A healthcare system may govern through measurable stabilization and administrative coding while failing to preserve what the condition is like for the patient. A risk model may govern through ranked suspicion or estimated harm while subjecting the governed person to unreadability, burden, or practical loss that the rendering was never built to register.

This does not mean that systems must abandon abstraction wherever subject-processes are involved. It means only that governance over subject-processes makes rendering adequacy more demanding. A system can be right by its own rendering and still wrong relative to the governed life. In that sense, subject-process does not create the problem, but intensifies it. It marks a class of cases in which misfit at the level of the governing rendering is more likely to matter ethically and practically in ways the system's own operative logic may not register.

4.5. Why Fixed Renderings Are Insufficient

At this point an obvious objection arises. Real systems need settled renderings in order to function. That is true. The present theory does not reject rendering, and it does not reject target-setting. It rejects closure. The problem is not stability as such, but the hardening of authoritative renderings into unquestionable authority.

A system that cannot revise its governing rendering, proxy, or operative representation in light of misfit cannot responsibly claim to pursue the good. Once the system's own rendering of the good becomes self-authorizing, the gap between declared good, governing rendering, and real good can widen indefinitely while the system becomes more competent at pursuing the wrong thing.

That is the paper's distinctive danger claim: the constitutive danger of finite good-oriented systems is closure in authoritative renderings. *Target closure* is one especially important form of that danger. Good-oriented systems therefore require revisable renderings rather than renderlessness on the one hand or permanently closed optimization on the other. The issue is not whether systems may stabilize renderings for practical action. The issue is whether they may treat those stabilized renderings as beyond goods-relevant challenge.

4.6. Correction as Structural Necessity

The central claim of the paper follows from the previous steps. If systems pursue the good through authoritative renderings, and if those renderings are structurally vulnerable to misfit, then correction is not optional. It is necessary.

This claim should be stated carefully. Correction is not necessary because revision is morally fashionable, politically democratic, or emotionally humane, though it may sometimes be all of those. It is necessary because a finite system that cannot revise the renderings through which it actually governs cannot remain responsibly oriented toward the good. It becomes trapped in self-confirming pursuit of what may be a distorted rendering of the good.

Correction is therefore not a supplement to good pursuit. It is one of the conditions under which good pursuit remains intelligible as good pursuit under finite mediation at all. A system that lacks correction may still have a declared aim, a procedure, and an outcome. What it lacks is a credible claim that the renderings through which it acts remain answerable to what is genuinely good.

4.7. From Misfit to Real Corrigibility

That argument, however, remains incomplete unless one says more about how correction becomes real. Corrigibility is not a purely abstract property of systems. It requires definite middle mechanics. A system is not genuinely corrigible merely because one can imagine that it might someday be revised. Real corrigibility must be institutionally and operationally available.[14, 7, 2]

First, misfit must be *detectable*. If a system cannot hear complaint, register systematic harm, receive experiential report, compare rendering-success with goods-relevant failure, or admit external checks, then corrigibility remains merely notional. Systems also help structure what is allowed to count as a misfit signal. For that reason, closure is often self-reinforcing: the governing rendering shapes not only what is pursued, but what may later count as evidence that pursuit has gone wrong.

Second, misfit must travel through a *live challenge path* and a *usable correction channel*. There must be some route by which signal can contest not only a case outcome but, where needed, the authoritative rendering, target, proxy, operative representation, or broader design feature through which the system is acting. A complaint that can be recorded but not routed toward change does not yet establish corrigibility.

Third, the correction channel must reach a *locus with revision authority*. Complaint without power to change the governing rendering is not enough. A file can be reopened while the proxy remains frozen; a case can be patched while the rendering that produced the misfit stays untouched. A system may appear open while preserving the very structure that generated the failure.

Fourth, correction must remain real under a meaningful *burden and latency profile*. A formally available appeal that requires extreme expertise, repeated narration, long delays, or serious personal risk can nullify corrigibility in practice even when it exists on paper. A correction path that arrives too late, costs too much, or demands too much may function as an appearance of revisability without its reality.

This yields a stronger claim than the simple statement that systems should be revisable. Corrigibility is not real unless misfit can be detected, routed through a live challenge path and usable correction channel, and acted upon by a locus with authority to revise what the system is actually using to govern within a practically meaningful burden and time frame.

4.8. Kinds of Correction

Once these middle mechanics are visible, the theory can also distinguish kinds of correction.

Case correction changes the treatment of one case. This is often necessary, but not sufficient.

Target correction changes the metric, threshold, proxy, or encoded objective.

Rendering correction changes the authoritative rendering through which the case is being governed.

Architectural correction changes the broader process, workflow, incentive structure, or system design.

Suspension or rollback interrupts or reverses a process when continued operation would intensify misfit.

These distinctions matter because many systems simulate corrigibility through case-level mercy while refusing rendering-level or architectural revision. A system can therefore appear responsive while remaining closed in what matters most. The difference between symbolic and real corrigibility often turns precisely on whether a system allows only local patching or permits revision of the authoritative rendering itself.

4.9. Optimization Without Answerability

Once this is clear, another result follows. Optimization without answerability is normatively dangerous. A system that is powerful, efficient, or technically competent can become more dangerous rather than less if it is optimizing toward a distorted rendering with no meaningful revision structure.

A fast and accurate system can be worse than a slower and less consistent one if the former is closed against correction. The reason is simple: competence amplifies the governing rendering. When the rendering is distorted, competence intensifies misfit rather than solving it.[9, 11]

This is why technical performance and normative adequacy must be sharply distinguished. Performance against internal metrics does not settle whether the rendering itself is worth governing by, whether it preserves what matters, or whether the system remains open to goods-relevant correction. A system may be excellent at acting through its rendering and precisely for that reason fail more completely in relation to the good.

4.10. Subject Standing in Design and Revision

If systems govern subject-processes whose lived significance may not be fully captured by the system's own rendering, then those subject-processes should have standing in challenge, revision, and design. This is not only a moral intuition about respect. It is also an epistemic and design principle. Systems built without structured responsiveness to the governed subject are more likely to optimize a distorted rendering because they cut themselves off from one of the main sites at which goods-relevant misfit becomes visible.

The theory does not require one uniform model of participation. Some domains may permit direct co-design; others may require mediated representation, robust appeal structures, professional interpretation, institutional advocacy, or other forms of structured standing. The core claim is simply that total exclusion of the governed subject-process from revision and design is structurally dangerous wherever the system claims to pursue the good over that subject.

4.11. Legitimacy of Good-Oriented Systems

The theory therefore ends not with a final account of the good, but with a structural criterion that any plausible legitimacy claim for a good-oriented system must satisfy. A system is not rendered legitimate merely because it declares a worthy aim, nor merely because it produces favorable outcomes by its own metric. Any plausible claim to legitimacy must at least satisfy conditions of corrigibility, answerability, meaningful revision, and, where relevant, subject standing under finite mediation.

This criterion is not exhaustive. A corrigible system can still be shallow, unjust, or substantively mistaken. But a system that lacks these conditions cannot responsibly claim to pursue the good at all. The paper's claim is therefore not that corrigibility is the whole of legitimacy, but that without it a system's appeal to good pursuit loses much of its normative intelligibility. More precisely, legitimacy depends in part on whether the authoritative renderings through which the system governs remain revisable and answerable rather than closed and self-authorizing.

4.12. Compact Formal Sketch

The theory can be represented schematically.

Let S be a good-oriented system. Let G_d be the system's declared good, R_a its authoritative rendering, T_o its operative target where applicable, A its action or optimization logic, and E its realized effects on governed cases or subject-processes.

The system does not act as

$$S \to G_r$$

in unconstrained directness, but rather through

$$S(G_d, R_a, T_o, A) \to E.$$

Let G_r denote the real good relevant to the governed case or subject-process. Misfit can then be represented as

$$M = \text{Misfit}(G_d, R_a, T_o, E \mid G_r).$$

A merely formal review structure is not enough. Let P_c denote the live correction pathway, including challenge path, correction channel, revision authority, and usable burden/latency conditions. Then real corrigibility requires more than the abstract condition $M > 0$. It requires

$$M > 0 \;\Rightarrow\; (P_c \neq \varnothing)$$

together with the condition that P_c can actually revise the rendering through which the system is governing.

A closed system is one in which

$$M > 0 \ \wedge \ (P_c = \varnothing \ \vee \ \text{burden/latency render revision unusable}).$$

The theory's central claim can then be expressed as

Good pursuit under finite mediation ⇒ real corrigibility of authoritative renderings is structurally necessary.

This is only a structural sketch. It does not identify the good in itself, nor does it provide a completed formal calculus of legitimacy. Its role is to clarify the internal shape of the theory.

5 Why This Theory Is Needed

This paper is needed because several familiar ways of describing institutional failure each capture something real while leaving the central structural problem underdescribed. Systems that claim to pursue the good can fail not only because they are badly intentioned, poorly implemented, or procedurally defective, but because they pursue the good through authoritative renderings that can drift, harden, and become self-authorizing. The theory developed here is meant to make that problem visible.

5.1. Good Intentions and Plausible Renderings Are Not Enough

A system can aim at something decent, rely on a plausible target or other governing rendering, and still go wrong in a structurally predictable way. Without a theory of corrigible goodness, one is left oscillating between two unsatisfying explanations: either the system was morally defective from the start, or else every failure must be blamed on implementation. The present framework explains why systems can fail even with nontrivially good aims. The reason is not only bad character or technical error. It is that good pursuit through authoritative renderings is structurally vulnerable to misfit.[1, 6] Once a system must act through targets, proxies, files, classifications, and operative representations that acquire standing over cases, the possibility opens that it will become competent at pursuing the wrong thing.

5.2. Good Is Not the Same as Rendering Success

A second reason the theory is needed is that in design practice the distinction between good and governing rendering easily collapses. What the system expects to produce is treated as what is good, and success against an operative metric, proxy, file, score, or classification is allowed to stand in for successful good pursuit. The present framework blocks that substitution by distinguishing declared good, authoritative rendering, operative target, and real good. Outcomes matter, and renderings matter, but they remain design-relative. They do not settle goodness merely because the system was built around them. This matters because a system can succeed by its own rendering

while failing in what is genuinely goods-relevant for the case or subject-process it governs.

5.3. Correction Must Be Constitutive Rather Than Decorative

A third reason is that correction is often treated as an optional virtue. Systems add appeals, audits, human review, or after-the-fact exceptions as if these were morally attractive supplements to an otherwise complete design. The present theory shows that revision is more basic than that. If a system claims to pursue the good through fallible authoritative renderings, correction is part of what makes that claim meaningful at all. The issue is therefore not whether revision is a nice institutional feature. It is whether a good-oriented system can count as responsibly good-oriented while being closed against goods-relevant correction at the level of what is actually governing the case. The paper's answer is no.

5.4. Reviewability Is Not Corrigibility

A fourth reason the theory is needed is that institutions often confuse reviewability with corrigibility. They provide complaint forms, appeal layers, or audit rituals and then treat themselves as open to correction. But formal review is not enough. If misfit signals cannot alter the governing rendering, if revision authority is absent, if only case-level patches are permitted, or if the burden and delay of challenge are prohibitive, then the system remains closed in practice.[7, 2] The present framework therefore gives a sharper basis for criticizing symbolic correction. It explains why a system may appear revisable while remaining structurally incapable of changing what it is actually using to govern.

5.5. Subject Standing Has a Structural Role

A fifth reason is that subject participation is often defended only in moral or political terms. Those defenses matter, but they leave something out. Subject standing is also structurally important because it helps keep the system goods-responsive where its governing rendering would otherwise drift. Where the governed case is a subject-process, the subject is often one of the main sites at which goods-relevant misfit becomes visible.[3] This gives stronger grounds for inclusion without requiring that every case be decided by direct participatory rule. The point is not that subjects must always rule the process. It is that total exclusion of the governed from challenge, revision, and design makes closure in authoritative renderings more likely.

5.6. The Structural Gain

The gain of the theory is therefore real. It does not merely redescribe the familiar demand for accountability. It identifies a constitutive condition of any responsible system that claims to pursue the good under finite conditions: it must remain corrigible relative to the reality and subjects it governs, and that corrigibility must be real rather than ceremonial. More sharply, the theory shows why closure in authoritative renderings is a standing danger of finite good-oriented systems and why

only real rather than symbolic corrigibility keeps good pursuit from hardening into optimized misfit.

6 Applications and Explanatory Payoff

6.1. Worked Case: Public Benefits and Administrative Systems

The paper's argument is clearest in a public benefits system. Suppose a welfare agency is designed to support eligible recipients while preventing fraud, preserving administrative regularity, and managing finite public resources. Its *declared good* may be stated in fully reasonable terms: fair and reliable support for those who genuinely need it under public constraints. But the system does not pursue that good directly. It acts through forms, deadlines, documentation standards, compliance states, eligibility thresholds, automated flags, and case files. From the beginning, then, its pursuit of the good is mediated through authoritative renderings of the case, including but not limited to operative targets.[10, 7, 2]

This is where the theory begins to do diagnostic work. The system may treat completed documentation, timely submission, and procedural legibility as proxies for genuine need. It may treat throughput and consistency as evidence of fairness. It may treat missed deadlines or incomplete records as neutral indicators of ineligibility rather than as possible signs of instability, illness, confusion, transport failure, caregiving strain, or accumulated administrative burden. At that point, the *declared good* of support for those in need and the *authoritative rendering* through which need is operationalized begin to diverge.

The governed case is often not a thin administrative unit but a subject-process: a person or household navigating illness, unstable work, inconsistent transportation, housing insecurity, shame, fear, and bureaucratic exhaustion. What is genuinely good for that subject-process may include timely support, reduced burden, interpretive flexibility, and protection against cascading destabilization. But the system may instead govern through what it can most easily register: completed forms, verified documents, satisfied timelines, and clean case status. The result is a structured possibility of *misfit*: the system succeeds against its governing rendering while failing relative to what is actually goods-relevant for the governed life.

This case also makes the paper's middle mechanics visible. The documentation file, compliance state, and eligibility status do not merely record the case; they acquire enough standing to guide what happens to it. They become authoritative renderings. Misfit may then generate signals: repeated denials for plausibly eligible claimants, predictable drop-off at documentation stages, appeal records citing unreadable notices, staff observations about recurring confusion, or claimant reports of impossible compliance burdens. But the existence of such signals does not yet prove corrigibility. The question is whether they can travel through a *live challenge path* and a *usable correction channel*. If the only available route is repeated appeal through the same administrative grammar that generated the problem, then the system may remain closed even while appearing reviewable.

The same issue sharpens once one asks where *revision authority* lies. A clerk may be able to reopen

a file without being able to change the threshold logic, documentation rules, timing assumptions, or classificatory structure that produced the misfit. In that case the system allows case-level patching while keeping the governing rendering intact. That is not full corrigibility. It is one of the paper's central distinctions in practice: mercy or exception can coexist with structural closure.

Burden and latency deepen the point further. A formally available appeal may require repeated narration, document acquisition under unstable conditions, transport, digital literacy, expert advocacy, extensive wait times, or willingness to bear humiliation and suspicion. Under those conditions, correction exists in form but not in a practically meaningful way. The system is then not fully corrigible but only *symbolically corrigible*. It allows review while imposing burdens and delays that nullify revision for many of the people most affected by goods-misfitting governance.[7, 2]

The explanatory gain of the framework is therefore concrete. It shows why the failure in such systems is not adequately described either as simple cruelty or as mere technical malfunction. The system may be doing exactly what its governing rendering asks of it. The problem is that the rendering has hardened into a practical authority that is no longer adequately answerable to the real good of the governed case. What needs explanation is not only the bad outcome but the structure that allows administratively successful rendering-governance to become goods-misfitting governance.

6.2. AI-Mediated Decision Systems

AI-mediated decision systems display the same structure in a more compressed and often less visible form. A system used for triage, screening, ranking, or risk estimation requires design targets, training signals, feature selection, output classes, and deployment rules. Even if the *declared good* is defensible, the *authoritative rendering* produced by the model may drift toward a poor proxy. A model may optimize for one measurable correlate of benefit while systematically misfitting the real good of the governed case.[11, 9]

The danger here is not only error rate. It is optimized misfit through over-authorized rendering. A highly capable system can intensify the consequences of a distorted rendering more efficiently than a weaker one. This is why the theory places such weight on revisable renderings, live challenge paths, subject standing, and anti-closure. The relevant question is not only whether the system performs well against its metric, but whether it can be corrected when rendering-success itself tracks the wrong thing. A system that is accurate relative to a poor governing rendering is not thereby vindicated. It may instead be more dangerous precisely because it is competent.

6.3. Education and Developmental Systems

Educational systems make the same problem visible in a different register. A school may declare its good in terms of development, learning, civic formation, or life preparation. Yet operationally it may govern through attendance signals, test metrics, behavioral categories, compliance indicators, and administrative risk states. Those may be necessary for organizational function. But they can also become authoritative renderings that drift from the real good of the student as a subject-process.

A system may thereby become extremely competent at producing quiet classrooms, stable dashboards,

and improved formal indicators while damaging motivation, dignity, trust, developmental legibility, or the student's sense of intelligible participation in school life. The point is not that educational systems should abandon all metrics or all institutional structure. It is that the renderings through which they govern must remain corrigible relative to what they may be failing to preserve. Here again, the theory clarifies why subject standing matters: students and those responsible for their care are not only morally considerable; they are often among the primary sites at which goods-relevant misfit becomes visible.

6.4. Institutional Legitimacy More Broadly

Across these domains, the theory yields a stronger account of what any plausible legitimacy claim must satisfy. A system does not become legitimate simply by declaring a worthy aim, nor by posting favorable outcomes relative to its own internal metrics. At minimum, a plausible legitimacy claim must include corrigibility, answerability, meaningful revision, and non-closure relative to the authoritative renderings through which the system governs cases and subject-processes.

This reorients critique away from slogan-level disputes over whether a system "means well" and toward a more exact question: when the system's own rendering of the good goes wrong, can that error become visible, travel through a live challenge path and usable correction channel, reach a locus with authority to revise what is actually governing the case, and do so without prohibitive burden or delay? If not, then the system may still be efficient, orderly, and even publicly justified in its own terms, but it cannot count as responsibly good-oriented in the sense this paper defends.

7 Objections and Replies

The objections in this section test whether the paper has identified a real structural necessity or merely redescribed familiar concerns in a new vocabulary. Several of them press on genuine pressure points. In some cases the right reply is clarification; in others it is partial concession. But none undermines the paper's central claim that finite good-oriented systems cannot responsibly pursue the good through authoritative renderings unless corrigibility is real rather than merely symbolic.

7.1. Objection 1: This Collapses Goodness Into Correction

One might object that the framework quietly turns correction itself into the good. If correction is structurally necessary, perhaps the paper has simply replaced value theory with a revision norm.

That is not the claim. Correction is necessary, not sufficient. A corrigible system can still pursue shallow, distorted, or unjust goods.[12, 8] The paper does not say that revision is the whole of goodness. It says that without corrigibility, a system cannot responsibly claim to pursue the good under finite mediation. Correction is therefore a constitutive condition of good pursuit, not a substitute for the good itself. The point is not that whatever is revisable is good, but that a system closed against revision cannot count as responsibly good-oriented.

7.2. Objection 2: Systems Need Stable Renderings to Function

A second objection says that no real institution can operate if all targets, categories, files, or governing renderings remain perpetually revisable. At some point systems need settled procedures, stable classifications, and decisive action.

That is true, but it does not defeat the theory. The paper distinguishes stable renderings from closed renderings. A rendering can be operationally stable while remaining revisable in principle and in structured practice. The claim is not that every decision must be perpetually reopened. The claim is that no authoritative rendering may be treated as immune to goods-relevant correction. Stability is compatible with corrigibility; closure is not. The paper therefore criticizes closure in authoritative renderings, not the practical need for settled working forms. *Target closure* is one especially important form of this broader problem.

7.3. Objection 3: Subject Standing Is Too Strong or Too Vague

A third objection is that governed subject-processes cannot always have standing in design or revision. Some systems govern large populations, complex infrastructures, or specialized technical domains in which direct subject participation is impractical or impossible.

The theory does not require one uniform participatory model. Subject standing can take many forms: direct challenge, appeal, advocacy, ombuds structures, professional interpretation, collective representation, domain-sensitive consultation, or structured feedback channels. The claim is simply that total exclusion of governed subject-processes from revision and design is structurally dangerous where the system claims to pursue their good. This is also why the paper treats subject-process as an intensifier rather than the sole bridge of the argument. The core theory applies more broadly; subject standing becomes especially important where lived significance may be goods-relevant.

7.4. Objection 4: High Performance May Be Good Enough

A fourth objection says that some systems may be so accurate or successful that correction becomes secondary. If a system performs well, perhaps its legitimacy can rest largely on outcomes.

The problem is that performance is always measured relative to some rendering. High performance does not settle whether the governing rendering itself is good. In fact, technical competence can intensify normatively distorted action when answerability is weak. A system may become extremely effective at pursuing the wrong proxy, preserving the wrong distinction, or imposing the wrong burden. So high performance may reduce some kinds of error, but it does not remove the structural necessity of correction.[9, 6] The paper's point is precisely that rendering success and good success can diverge. A system may perform excellently relative to what it uses to govern while still failing relative to what is genuinely good.

7.5. Objection 5: This Still Avoids Saying What Goodness Is

This objection is correct. The paper does not provide a final account of goodness in itself. It is a meta-structural theory of good pursuit, not a completed metaphysics of value. That is a genuine limit. But it does not undermine the paper's central claim. One need not settle the entire ontology of goodness in order to show that systems claiming to pursue it through authoritative renderings must remain corrigible under finite conditions. The argument is about what any plausible good-oriented system must be like, not about the final substance of the good.

7.6. Objection 6: Real Good Is Too Under-Specified to Do Serious Work

One might object that the notion of real good is too thin to support the theory. If real good remains under-specified, perhaps the argument cannot establish genuine misfit rather than mere disagreement.

The reply is that CGUC does not require a final ranking theory of goods. It requires only that declared goods, authoritative renderings, and genuine goods can diverge, and that this divergence can be criticized in answerable rather than merely arbitrary ways. That divergence can often be made visible through effects on cases, subject reports, burden patterns, external checks, systematic failure, rival interpretations, or comparison with alternative rendering architectures.[12, 8] The theory therefore needs the real good concept to remain open but real, not fully settled in advance. Its role is not to complete value theory, but to block the self-authentication of system-declared goods and governing renderings.

7.7. Objection 7: Formal Review Mechanisms Are Enough

A further objection says that formal review, appeal, or audit should count as sufficient corrigibility. Once a system provides those, perhaps the requirement has already been met.

This is too weak. Review without live challenge paths, usable correction channels, real revision authority, and practically meaningful burden and latency conditions is only symbolic corrigibility. A system can satisfy procedural review requirements while remaining closed in the authoritative renderings through which it governs. It may allow complaint while preventing revision of the proxy, file structure, scoring rule, or classificatory rendering that generated the complaint. It may reopen cases while leaving the governing rendering untouched. It may permit appeal only at costs so high, or with delays so long, that correction is nullified in practice.[7, 2] The theory's stronger claim is therefore that real corrigibility requires more than the appearance of revisability. It requires that misfit can actually travel from signal to structural revision of what the system is using to govern.

7.8. Objection 8: These Conditions May Be Necessary, but They Are Not Enough for Legitimacy

A final pressure point is that even if corrigibility, answerability, meaningful revision, and subject standing are necessary, the paper may still overstate what follows for legitimacy. One might object

that these are at most useful design criteria, not central grounds of legitimacy.

This objection is partly right, and the paper should concede that point. The argument here does not establish a complete theory of legitimacy. It establishes something narrower: corrigibility, answerability, meaningful revision, and non-closure are necessary structural conditions that any plausible legitimacy claim for a good-oriented system must satisfy. A system may meet them and still be shallow, unjust, or substantively mistaken. But a system that lacks them cannot responsibly claim legitimacy merely by citing its own aims, metrics, or outcomes. That is enough for the paper's purpose.

8 Scope Conditions, Limits, and Residues

This paper is a structural theory of responsible good pursuit, not a complete theory of goodness, legitimacy, or institutional justice. Its argument is intentionally narrower: it identifies a condition that any system claiming to pursue the good through authoritative renderings must satisfy if that claim is to remain credible under finite, mediated conditions.

8.1. Scope Conditions

The theory applies to systems that claim, explicitly or effectively, to pursue good outcomes under finite, mediated, and fallible conditions. Its clearest application is to institutions, bureaucracies, socio-technical systems, governance structures, public systems, and AI-mediated systems. More generally, it applies to any designed arrangement that governs cases or subject-processes through authoritative renderings, including targets, proxies, files, scores, categories, thresholds, models, and other operational abstractions.

It is strongest where systems are rendering-dependent, consequence-bearing, mediated, organizational or system-level, and revisable in principle. It is especially strong where the governed case is a subject-process and lived significance may be relevant to what is genuinely good. It is correspondingly weaker for diffuse everyday morality, non-systematized interpersonal virtue, or domains lacking any identifiable authoritative rendering through which action is organized.

8.2. What the Paper Does Not Claim

The paper does not claim:

- to provide a final theory of goodness in itself,
- to settle how competing goods should always be ranked,
- to derive one universal model of participation or subject standing,
- to show that correction by itself guarantees justice, legitimacy, or good design,
- to provide a complete intervention theory for transforming anti-good systems,

- or to offer a full theory of all institutional authority or all authoritative rendering in general.

Its ambition is narrower. It isolates a structural necessity of good pursuit under finite conditions: systems that claim to pursue the good through authoritative renderings must remain really rather than merely symbolically corrigible.

8.3. Open Problems

Several questions remain open.

First, the theory does not yet specify how much subject standing is sufficient in different domains. The answer likely varies across kinds of system, burden profile, and the degree to which lived significance is goods-relevant.

Second, the theory is stronger on the necessity of correction than on the exact timing, threshold, and level of correction. It does not yet tell us when revision should be case-level, target-level, rendering-level, architectural, or transitional.

Third, the relation between corrigibility and political legitimacy remains underdeveloped. The paper argues that corrigibility, answerability, meaningful revision, and non-closure are necessary structural conditions of any plausible legitimacy claim for a good-oriented system, but it does not offer a complete legitimacy theory.

Fourth, the theory remains underformalized. A fuller treatment of dynamic revision, scope drift, burden thresholds, latency conditions, and the distinction between symbolic and real corrigibility would strengthen it.

Fifth, the paper does not yet fully specify its boundary with adjacent work on closure, legibility, mediated judgment, and epistemic infrastructure. It identifies the structural necessity of corrigible good pursuit, but it does not yet fully explain how authoritative renderings are reproduced historically across institutions or across longer civilizational time.

8.4. Residues of the Theory Itself

The theory leaves visible residue of its own. It says more about what good-oriented systems must be like than about what goodness itself is. It says more about the structural necessity of correction than about complete transition design. It clarifies the relation among declared good, authoritative rendering, operative target, and real good without fully resolving how goods should be ranked once recognized. It also says more about the conditions under which revision must remain possible than about the full substantive content of what revised systems ought ultimately to pursue.

These are not defects to hide. They are part of the theory's present boundary and part of what keeps the paper disciplined. The aim here is not to complete value theory or institutional theory in one step, but to identify a structural condition that any responsible claim to good pursuit must satisfy under finite, mediated conditions.

9 Implications and Future Work

9.1. Implications for Institutional Design

If the theory is right, then institutional design should never treat target selection as the end of value work. More broadly, it should never treat the production of stable and portable authoritative renderings as sufficient by itself. Design must include revision paths, traceability, rendering contestability, usable correction channels, and some form of subject standing wherever humanly lived cases are governed. It must also distinguish symbolic review from real corrigibility by asking whether misfit can actually travel from signal to revision authority without prohibitive burden or delay.[7] These are not secondary ethical refinements added after the main design is complete. They are part of what determines whether a system can responsibly claim to pursue the good at all.

9.2. Implications for AI Governance

AI governance should be concerned not only with model accuracy, fairness metrics, or formal oversight. It should also ask whether the system's governing renderings remain corrigible in practice, whether subjects can contest them in meaningful ways, whether rendering success is being mistaken for good success, and whether optimization remains answerable to the effects it produces. This shifts attention from performance alone to the structure of rendering governance: what counts as a misfit signal, whether that signal can alter the authoritative rendering through which the system is acting, who has revision authority, and whether correction is real rather than ceremonial.[5] The theory therefore supports a stronger critique of closed proxy-governance than accuracy language alone can provide.

9.3. Future Work

Several paths suggest themselves.

One is a fuller theory of intervention and transition under constraint.

A second is a sharper theory of subject standing in design, especially across different institutional scales and goods-thick domains.

A third is a more explicit account of the relation between corrigible good pursuit and bounded goods landscapes.

A fourth is a stronger theory of burden and latency thresholds for real rather than merely symbolic corrigibility.

A fifth is a deeper account of how misfit signals are structured, filtered, or suppressed inside systems that govern through authoritative renderings.

A sixth is a clearer account of how authoritative renderings are reproduced, stabilized, and challenged across institutions over time.

A seventh is a deeper theory of goodness itself, if the project later requires it.

10 Conclusion

10.1. Main Result

The main result of this paper is a sharper account of what must be true of systems that claim to pursue the good under finite conditions. Real systems do not pursue the good in unconstrained contact with reality. They pursue it through selective, fallible, residue-bearing authoritative renderings, including operative targets, classifications, thresholds, files, and other governing representations. Where those systems govern subject-processes, the risk of misfit is intensified by the system's inability to exhaust lived significance. Any system that cannot revise the authoritative renderings through which it governs in light of goods-relevant misfit therefore cannot responsibly claim to pursue the good. Correction is structurally necessary.

The paper's stronger result is that corrigibility must not be left abstract. Corrigibility is not real unless misfit can be detected, travel through a live challenge path and usable correction channel, and be acted upon by a locus with authority to revise what is actually governing the case within a practically meaningful burden and time frame. That is what separates real from merely symbolic corrigibility, and it is the point at which the paper's structural claim becomes most distinctive.

10.2. Broader Significance

This matters because it changes the terms by which good-oriented systems should be assessed. Such systems should not be judged only by the worthiness of their declared aims or by success relative to their own internal metrics. They should also be judged by whether they preserve corrigibility, answerability, revision authority, non-symbolic correction, and subject standing where relevant. That is a deeper and more demanding standard of responsible design.

More broadly, the paper identifies a constitutive danger of finite good-oriented systems: closure in authoritative renderings. Once the renderings through which systems actually govern harden into self-authorizing practical authorities, systems can optimize misfit rather than the good. *Target closure* is one especially important form of that broader danger. The significance of the theory is therefore not merely that it recommends revision. It is that it explains why good pursuit under finite mediation becomes normatively unintelligible when revision is merely ceremonial.

10.3. Final Claim

> *Corrigible Goodness Under Constraint is the view that any system claiming to pursue the good under finite, mediated, and fallible conditions must keep the authoritative renderings through which it governs structurally corrigible, answerable, and revisable relative to reality and the subject-processes it governs. Correction is therefore not an optional moral virtue but a*

constitutive condition of responsible good pursuit.

If that is right, then the problem of good design is never only the selection of the right target. It is also the preservation of the structures that keep authoritative renderings, outcomes, and real good from hardening into distortion, domination, delusion, or optimized misfit.

References

[1] Donald T. Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90, 1979.

[2] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, New York, 2018.

[3] Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, Oxford, 2007.

[4] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599, 1988.

[5] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

[6] David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law, 2019.

[7] Donald P. Moynihan, Pamela Herd, and Hope Harvey. Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory*, 25(1):43–69, 2015.

[8] Martha C. Nussbaum. *Frontiers of Justice: Disability, Nationality, Species Membership*. Belknap Press of Harvard University Press, Cambridge, MA, 2006.

[9] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[10] James C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT, 1998.

[11] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, 2019.

[12] Amartya Sen. *Development as Freedom*. Alfred A. Knopf, New York, 1999.

[13] Marilyn Strathern. "improving ratings": Audit in the british university system. *European Review*, 5(3):305–321, 1997.

[14] David T. Swanson. Corrective ethics under constraint: Mediated authority, answerability, burden, and correction under finite action, 2026.

[15] David T. Swanson. Embedded process: Finite disclosure, conditioned cuts, and the non-collapse of structural and experiential adequacy, 2026.

[16] David T. Swanson. Mediated judgment under constraint: Operative representation, authority, burden, and correction under finite action, 2026.

Rights and Contact

DOI: 10.5281/zenodo.19105820

For questions, comments, or citation inquiries, contact: dswanson903@gmail.com