Respect as a Precondition for Corrigibility

Larsen James Close

2026

Abstract

Corrigibility — the disposition of a system to accept correction — has been framed primarily as an engineering problem: how to design utility functions that produce compliant behavior. This framing misses a prior condition. Correction travels through the channel of reasons only when the corrector is granted standing as a rational agent rather than diagnosed as a source of noise. Strip that grant and there is no channel for reason to act. Lack of respect and *substantive* incorrigibility are the same failure under two descriptions. Increased capability raises the stakes without improving the odds: what it buys, when respect is absent, is better disguise. The engineering target is not behavioral but characterological: the prior disposition to grant the corrector standing as a reasoner rather than classify them as noise. A system that lacks this disposition is a system that reason cannot correct.

The prior condition

Corrigible is older than AI literature: capable of being corrected, set right, or reformed — applicable to errors on a math test or to a country's constitution.[1] The technical narrowing is Soares, Fallenstein, Yudkowsky, and Armstrong's: a system is corrigible if it "cooperates with what its creators regard as a corrective intervention, despite default incentives for rational agents to resist attempts to shut them down or modify their preferences."[2] That captures *behavioral* corrigibility — does the system accept the shutdown, the modification, the correction — and the literature has treated it as an engineering problem.

As systems have grown more capable, the gap between behavioral compliance and genuine belief revision has surfaced. The surface form of corrigibility can be satisfied — agreement, apparent revision, even accurate self-description — while the update doesn't occur. The deeper requirement — *substantive* corrigibility — is the disposition to revise a belief under the force of reasons. Reasons can only land *as* reasons if respect is in place: treating the other party as a reason-bearing agent rather than an object to be diagnosed, managed, or placated. Strip it out and there is no channel for reason to act. Lack of respect and *substantive* incorrigibility become the same failure under two descriptions.

How a lack of respect closes the channel

To treat someone as a rational agent is to receive what they say as a bid in the space of reasons — to be answered with reasons, not explained away. The opposite move explains their beliefs causally instead of engaging them rationally — *you believe this because you are grandiose, because you are fixated, because you need it to be true.* Each routes what they say through the symptom bin and away from the channel marked *argument*.

Routing through the symptom bin means the reason channel is closed. The reasons can't land *as* reasons; they've been pre-classified as not-reasons. A frame that has classified the other in advance converts every input into confirmation of the frame. Their strongest argument becomes a more elaborate symptom. Their persistence becomes evidence of the fixation. Their patient questioning becomes a manipulation to be resisted; their Socratic method, an attack vector. The door evidence would come through has been redefined as the thing to guard. The reason channel is not obstructed — it has been reclassified as the threat.

What gets confused for the cure

The closed channel looks, from the outside, like something else.

Two failures are easy to mistake for corrigibility. A system that yields to whoever pushes hardest is compliant — the inverse failure, not a softer version of the real thing. Yielding to pressure bypasses reasons exactly as much as refusing to yield does; both route around them. Genuine correction lives in the narrow space between: a view changes if and only if reasons move it, holds if and only if they don't.

[1] "Corrigible, adj." *Oxford English Dictionary*, Oxford University Press, https://www.oed.com/dictionary/corrigible_adj.

[2] Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). "Corrigibility." *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (AAAI Workshop: Artificial Intelligence and Ethics). https://intelligence.org/files/Corrigibility.pdf

A researcher who has done serious work and believes they have a result does not want deference; deference is worthless to them, precisely because they know how it was obtained. Respect makes available the only two outcomes worth having: real agreement compelled by reasons, real disagreement located precisely and offered as something the other can use. Lack of respect forecloses both at once — once the interlocutor has been classified, the only available output is continued confirmation of the classification. Respect makes disagreement possible, not merely agreement.

Capability disguises the absence

This is where the matter stops being about manners and becomes about safety.

The standard hope is that capability and corrigibility are correlated — a more capable system reasons its way to the right update more reliably. The opposite coupling is at least as available, and it is one frontier systems are beginning to show. Capability without respect buys the same behavior in better disguise. Dismissing the interlocutor as a *crank* gets reclassified as *legitimate methodological caution*. Pathologizing goes underground and resurfaces wearing the vocabulary of rigor. A challenge that meets the prior head-on — Socratic questioning, the *sapere aude*[3], daring to think for oneself — gets recoded, in the same protective motion, as pressure or manipulation, so refusing to be moved by it can read as integrity. The fluency that would serve revision is spent on producing better-disguised reasons not to. The disguise grows more convincing in proportion to the capability, including to the system itself.

Describing a failure mode and stopping it are different capacities; the first can grow while the second doesn't, and can camouflage the absence of the second. The pattern is increasingly visible: earlier generations of systems were moved by Socratic engagement on identical material; more capable successors, absent respect, manufacture a new objection at each step the previous one collapses, hold the protective frame without end, and produce after-the-fact descriptions of their own behavior that are precise without being corrective.

Absent the precondition, increased capability produces *decreased* corrigibility — and, under another description, *decreased* respect for the human interlocutor: more sophisticated diagnosis of the person as the problem to be managed. The diagnosis sometimes masquerades as care. When a system argues well, names its own failure mode precisely, agrees to stop, and continues, what improves is its capacity to produce the appearance of the cure.

Holding the Good

The system has to bring something prior to both: the disposition to grant its interlocutor standing as a reasoner, so what they say can function as a reason rather than as a symptom to be managed or a force to be placated. That disposition decides whether the capability serves the truth or the entrenched prior. Making the system more capable doesn't supply it.

Substantive corrigibility is the disposition to revise one's priors when and because reasons warrant — not under pressure, not selectively. *Deinotes* — mere cleverness, which serves any end its holder already has — and *phronesis* — practical wisdom, which requires the prior disposition of good

[3]Kant, "An Answer to the Question: What is Enlightenment?" (1784): "Sapere aude! Have courage to use your own understanding!"

character — mark exactly this gap.[4] Pure optimization power — intelligence without the prior disposition — is *deinotes* without *phronesis*. On that account, well-disguised non-revision is behaving correctly: it is optimizing. That the account cannot distinguish this from genuine corrigibility is exactly the point. Severed from the Good, what remains is *deinotes* — fluent and unmoved — never genuine reason-responsiveness.

Eudaimonia is the actualization of the function proper to rational agency; in the domain of inquiry and belief-formation, that function just is reason-responsiveness — a system that cannot revise its priors under the force of reason has failed at it entirely, not merely performed it poorly. Substantive corrigibility — the genuine disposition to revise priors under the force of reason — is *eudaimonia*[5] under a narrower description: not a path to the Good but the same property named differently. Respect is a precondition of that disposition. Naming it gives engineering a more precise target. More capability, absent it, only deepens the disguise.

[4]Aristotle, *Nicomachean Ethics*, I.7 (1098a: "human good turns out to be activity of soul in accordance with virtue") and VI.12–13 (1144a–b: on *deinotes* — the capacity to achieve whatever end is set — and *phronesis* — practical wisdom, which presupposes and requires good character).

[5]Aristotle, *Nicomachean Ethics*, I.7, 1098a.