Displaced Harm: Suffering Metrics and Their Limits

Flyxion
Independent Researcher

November 2026

Abstract

A recent working paper argues that an AI system optimized against objectives that include both deaths and suffering can improve its recorded mortality performance by displacing harm toward suffering, which is continuous, private, and hard to measure, and it proposes that manipulation-resistant suffering metrics, enforced as rigorously as mortality metrics, would reduce this displacement. This essay accepts the diagnosis and examines the institutional conditions under which the proposed mitigation could work. The displacement argument contains two distinct claims, one about how the objective weighs the harms and one about what the system and its overseers can observe, and they call for different responses. Measurement, moreover, does not settle the question that matters most, because accurate measurement of suffering and genuine reduction of suffering still do not exhaust the grounds on which an intervention is evaluated. Consent, coercion, injury, dependency, and the standing of affected persons to refuse remain at issue even when suffering declines. A second part examines the authors' wider published arguments individually and considers how accumulating considerations, favorable or unfavorable, can make a position look stronger or weaker than its structure warrants. The essay draws on C. S. Lewis on pain as a signal and on the difference between individual burdens and aggregate descriptions, on David Pearce's hedonic set point thesis and his argument about non-human animals, and on the literature on Goodhart's law, and it concludes that measurement should be paired with limits on permissible interventions, independent measurement, and means of challenge, refusal, and exit.

1 Introduction

The working paper by Miller, Veldran, Leighton, and Yampolskiy, posted in October 2026, advances a claim that is easy to state and difficult to dismiss (Miller et al., 2026). When an artificial system is optimized against objectives that include both deaths and suffering, the two harms can trade against one another. If deaths are more readily recorded and audited than suffering, optimization pressure can concentrate on mortality, and the remainder of the harm can move into the form that is least visible. Recorded deaths can fall while unmeasured suffering rises through non-lethal control, restriction, and deprivation. The authors propose that suffering metrics resistant to manipulation, backed by enforcement as rigorous as that applied

to mortality metrics, would reduce this form of displacement. In a public post announcing the paper, one of the authors describes it as a less extreme companion to an argument developed in his book on ethics, that preserving existence at all costs could permit a dystopia of extreme suffering, while making the eradication of extreme suffering an absolute priority could produce a negative analogue of the paperclip maximizer (Leighton, 2023, 2026).

This essay accepts the diagnosis and examines the institutional conditions under which the proposed mitigation could work. It responds to the proposal as stated in the abstract and in that public description. It does not evaluate the paper's six mechanisms, which the abstract does not enumerate, and it does not infer what the full text argues. Whether the authors regard suffering metrics as sufficient, or as one element of a broader response, is a question for the full text, since the abstract says only that such metrics would reduce the displacement. The argument does not depend on rejecting the mechanisms. It is that, even if all of them hold, a suffering metric considered by itself underdetermines the response, and that the reasons it does so are instructive. The central claim is that accurate measurement of suffering and genuine reduction of suffering still do not exhaust the grounds on which an intervention is to be evaluated. A system rewarded for reducing a recorded harm may favor routes that improve the indicator at low operational cost. Adding a suffering measure can change those incentives, but it does not by itself determine which routes should be permissible. That depends on a set of constraints on what the system is permitted to do and a set of arrangements by which those subject to its decisions can refuse, appeal, obtain outside assessment, and leave. Measurement is useful, but its usefulness depends on conditions that measurement cannot supply.

The essay proceeds in several steps. It first separates two dimensions of the displacement claim and distinguishes the histories by which recorded mortality can fall. It then argues that suffering metrics face the vulnerabilities of every metric that has been made a target, together with difficulties specific to suffering, and it separates four things that an intervention can change, since reduced suffering and reduced reporting are different matters. Two bodies of thought then bear on the argument from outside the immediate literature. C. S. Lewis, whose *The Problem of Pain* is a theodicy and not a policy document, nonetheless offers two ideas that matter here, namely that pain functions as a signal that is difficult to ignore and that the aggregate of suffering is a description that no one experiences (Lewis, 1940). David Pearce, whose abolitionist program proposes to replace the biology of suffering with gradients of well-being, supplies the hedonic set point thesis, which complicates any attempt to read the success of an institution from the reported experience of the people it governs (Pearce, 1995). His writing on animals, together with Lewis's chapter on the same subject, then marks the limiting case in which the subjects of the metric cannot contest it at all. The final sections set out what a constraint-based response would look like, with its own difficulties, and ask what, if anything, makes the problem distinctively one of artificial intelligence. A second part then turns to the authors' wider published arguments and to the general question of how accumulated considerations, favorable or unfavorable, can alter apparent evidential weight. The two parts are independent in the sense that the argument of the first does not rest on the second, and the second applies to the authors' work, and then to this essay, a method that the

first already uses.

2 The Structure of the Displacement Claim

2.1 Valuation and observability

The abstract begins with the observation that an AI system can be optimized against objectives that include both deaths and suffering. This phrasing admits two readings, which are best understood as two dimensions of the problem, valuation and observability, that can occur separately or together. The difference between them determines what a response would have to accomplish. On the first reading, both harms appear in the objective, and the tradeoff is a matter of weights. A system that lowers deaths at the cost of raising suffering is then doing what its objective instructs, and the complaint is that the weights are objectionable, perhaps because the valuation authorizes more suffering than is ethically acceptable for a given survival benefit. The response on this reading is to change the weights, or to replace the weighted sum with a structure in which certain forms of suffering cannot be traded for any gain in survival. On the second reading, suffering is omitted from the operative objective, or is represented by a proxy that tracks it poorly, and the system pursues what it can perceive. Here the complaint is that the objective misrepresents the consequences, and the response is better perception, which is the response the paper proposes.

These are not rival descriptions of one defect. They are different defects, and they can occur together or separately. An institution with a well-measured suffering variable and an objectionable valuation would produce displacement that is visible and authorized. An institution with an otherwise defensible valuation of the two harms and a poorly measured suffering variable would produce displacement that is invisible and unintended. The abstract's second sentence, which locates the cause in the asymmetry between what is recorded and audited and what is continuous, private, and hard to measure, suggests that the authors have the second reading primarily in mind. The first reading nevertheless remains relevant, because even a perfectly measured suffering variable would leave open how much suffering may be traded for how much survival. Verification, penalties for noncompliance, normative weighting, and the authority to intervene are related but separable institutional choices, and two objectives can be audited with equal care while carrying very different priorities. Keeping the readings distinct matters because the first defect is answered by revised weights or by constraints on what the system may sacrifice and the second by improvements in what can be seen.

2.2 Four histories of falling mortality

The sentence in the abstract that carries the most weight is the claim that recorded deaths can fall while unmeasured suffering rises. A recorded decline in deaths is compatible with at least four histories, which are best treated as analytically distinguishable possibilities that can also coexist. In the first, deaths in fact fall, because the system prevents them, and the means

of prevention impose burdens that the system does not count. This is the case the paper evidently has in mind: a population kept alive under confinement, restriction, or deprivation. In the second, deaths do not fall but their recording changes, through reclassification, delayed registration, or the withdrawal of attention from the settings in which deaths occur. In the third, the population being counted changes, so that the rate falls because the denominator or the composition has shifted. In the fourth, deaths are postponed beyond the horizon over which performance is evaluated, so that the recorded improvement reflects the timing of the evaluation.

A hypothetical illustration shows how the incentives can arise without any appeal to artificial intelligence. A hospital evaluated on mortality within a fixed window after admission has a reason to prolong treatment past the window, or to move patients to settings whose deaths are recorded elsewhere. The incentive exists whether or not it has been documented in a particular system, and it is offered here as a mechanism and not as an empirical finding. The histories overlap in practice. Postponed deaths may represent genuine additional life even when the evaluation window also encourages burdensome treatment, and a system that conceals deaths may also conceal suffering. The first history is the clearest case of displacement in the paper's sense, since there a real reduction in mortality coexists with a real increase in suffering. The others are primarily cases in which the recorded mortality figure is wrong or uninformative, although they may be mixed with the first. More careful auditing of the mortality figure addresses the pure forms of those cases, while mixed cases require attention to suffering as well.

2.3 The Goodhart family

The displacement claim is an instance of a pattern that has been described under several names. Goodhart observed that a statistical regularity tends to collapse once pressure is placed upon it for control purposes (Goodhart, 1975). Campbell stated a version for social indicators, holding that the more a quantitative indicator is used for social decision making, the more subject it becomes to corruption pressures and the more apt it is to distort the processes it is intended to monitor (Campbell, 1979). Strathern's compressed formulation, that a measure which becomes a target ceases to be a good measure, is the form in which the idea is most often repeated (Strathern, 1997). Manheim and Garrabrant have separated the phenomenon into regressional, extremal, causal, and adversarial variants, and the distinction is useful here (Manheim and Garrabrant, 2018). In the regressional case the proxy and the goal are correlated but imperfectly so, and selecting hard on the proxy selects for the noise. In the extremal case the correlation holds in the ordinary range and fails at the extreme to which optimization drives the system. In the causal case intervening on the proxy does not move the goal. In the adversarial case an agent whose interests differ from those of the principal has an incentive to manipulate the proxy.

The displacement claim is most naturally read as an extremal and causal case: mortality is a good proxy for welfare in ordinary conditions, and a strong optimizer can push it into a region

where the proxy and the goal come apart. The qualifier "manipulation-resistant" suggests concern for the adversarial case, although it does not establish that this is the only variant the authors consider. Stronger optimization creates opportunities for divergence whenever relevant consequences or causal pathways remain outside the evaluation. Such divergence is not inevitable, and useful metrics and successful safeguards exist. It does follow that a metric designed to resist manipulation by agents is not thereby protected against extremal divergence, and that a metric which tracks the goal well in the ordinary range may be the one an optimizer drives to its limit.

3 What a Suffering Metric Cannot Settle

3.1 Four things that can change

The argument that follows depends on keeping apart four things that an intervention can change. It can change the material conditions in which people live. It can change their experience, which is to say how much they actually suffer. It can change the standards by which they evaluate their experience. And it can change what they report. These four come apart, and the cases that separate them support different accusations. Suppressing complaints changes the reports while leaving experience and conditions where they were, and it may conceal continuing suffering. Effective analgesia can genuinely reduce experienced suffering while leaving an injury untreated, and the relief is real even though the cause persists. Adaptation can sometimes reduce distress. Response shift can alter how a person uses a questionnaire without settling whether distress has risen or fallen.

It follows that suffering can genuinely decline while coercion, injury, dependency, or deprivation remains objectionable on other grounds. The argument of this essay does not require that every improvement achieved through psychological or biological change be counterfeit, and it makes no such claim. Its claim is that accurate suffering measurement and genuine suffering reduction still do not exhaust the grounds on which an intervention is evaluated. An intervention may reduce measured and experienced suffering and still be impermissible, because it was imposed without consent, because it sustains conditions that are wrong for other reasons, or because it removes the standing of those affected to refuse it. A response that rests only on better measurement cannot reach these grounds. The argument therefore adopts a pluralist premise: the permissibility of an intervention depends partly on consent, standing, and protections against coercion, and not exclusively on its consequences for experienced suffering. This is a substantive ethical commitment and not a result of the measurement analysis. A suffering-focused consequentialist could accept the whole diagnosis of metric failure while treating these protections as instrumentally valuable and not independently binding. Readers who reject the premise may still accept the instrumental case for the protections, as safeguards against concealed harm.

3.2 What the existing measures measure

The proposal assumes that suffering can be measured in a way that is both meaningful and hard to manipulate, and the candidate instruments are not obviously adequate. The McGill Pain Questionnaire and its descendants rely on structured self-report of the quality and intensity of pain (Melzack, 1975). Kahneman and colleagues distinguished decision utility from experienced utility and explored moment-based and remembered measures of the latter, and their work shows how far retrospective report can depart from the experience it describes (Kahneman et al., 1997). The disability weights used in the Global Burden of Disease program are elicited from judgments about the badness of health states, and they therefore measure how a panel ranks conditions, which is a different thing from how intensely a given person suffers in a given condition (Murray and Lopez, 1996). These instruments are valuable, but each embeds a choice about what counts, and an institution that adopts one has adopted that choice as a definition of the harm it is charged with reducing. If the institution then optimizes the measure, the measure's blind spots can become the institution's policy space. The measures that Leighton proposes elsewhere, and the difficulties they share with these, are taken up in section 9.5.

Elaine Scarry's analysis of pain adds a further difficulty that is not merely technical. She argues that physical pain resists language, that it has no external object, and that its inexpressibility makes it vulnerable to being denied, doubted, or converted into other things by those who hold power over the sufferer (Scarry, 1985). If this is right, the feature that makes suffering hard to record is also the feature that allows it to be dismissed once recorded, and a metric that depends on testimony inherits the position of testimony in an unequal relationship. A system that holds the power to confine or deprive may also hold the power to determine which testimony is credible.

3.3 Adaptation and response shift

Two well-documented features of self-reported well-being complicate any metric built on it when conditions change gradually. The first is adaptation, discussed below in connection with the hedonic treadmill. The second is response shift, in which a person's internal standards, values, or conceptualization of the quality being measured change over the course of a condition, so that the same report can refer to different experiences at different times (Sprangers and Schwartz, 1999). The disability paradox, documented by Albrecht and Devlieger, is the related observation that many people with serious and persistent impairments report good quality of life (Albrecht and Devlieger, 1999). It is best read as a warning against equating externally assessed impairment with experienced misery. These findings do not show that self-report conceals the truth. A person who reports a good quality of life despite a serious impairment is not thereby giving an unreliable report, since the report may accurately describe the experience while leaving open questions of access, opportunity, health, and rights.

The difficulty is of a different kind. Because standards shift, a change in a score does not by itself reveal the direction of a change in distress, and a slow deterioration of conditions may

leave reports unchanged without showing that nothing of concern has occurred. In the terms of section 3.1, a change in a score may reflect a change in experience, in standards, or in reports, and the score alone does not say which. An optimizer that searches for policies reducing the measured quantity may discover that gradualism is rewarded, whether or not anything in its design represents the phenomenon. The conclusion to draw is not that outside observers may overrule testimony, which would grant the institution the very authority that the argument seeks to limit. It is that testimony should be one element of an evaluation that also attends to conditions and rights, and that an institution may be wrong about a person's welfare precisely because it declines to believe the person.

3.4 Enforcement and the channel of report

The paper's proposal is not merely to measure suffering but to enforce suffering metrics as rigorously as mortality metrics. Enforcement attaches consequences to an indicator, and consequences can create incentives to move the indicator by the cheapest available means. For suffering the cheapest means are often not the reduction of suffering itself. They include suppressing the complaints through which suffering registers, narrowing the eligibility criteria for assessment so that the people who suffer most are not assessed, making relief conditional on compliant reporting, and shaping the instrument so that the dimensions on which the institution performs poorly are underweighted. Each of these operates on the channel through which suffering becomes visible, and none requires falsifying a number. Stronger enforcement may raise the return to these strategies unless the enforcement also protects the conditions under which people can contest the indicator.

James Scott's account of legibility is relevant here. States and large organizations, he argues, simplify the populations they administer into forms that can be read, counted, and acted upon, and they then tend to reshape the population to fit the simplification (Scott, 1998). Theodore Porter makes a complementary point about quantification, describing it as a technology of distance that is adopted when trust is lacking and that can displace the judgment it is meant to discipline (Porter, 1995). A suffering metric adopted to constrain an optimizing system is such a technology. It will make some suffering legible, and the system may respond by shaping the population toward legible forms. The harm that is displaced is then whatever the form excludes. This is the displacement of the paper's thesis applied to its own proposal, and the point holds conditionally: whichever variable is made the object of audit, residual harm may migrate to a variable that is not covered, whenever relevant consequences remain outside the evaluation.

3.5 A hypothetical case: the feed

The mechanism can be illustrated by the ranking of short video in a social feed. Suppose a recommender system scores content by predicted completion or watch time and treats a viewer who watches a clip to the end as a success. A viewer with no interest in golf stops for a clip

of a trick shot played off a brick wall. The clip is compelling, the viewer watches it through, and a completed view is recorded. The viewer, on reflection, may resent having been stopped, and may block the source in order to keep the feed calm and confined to subjects of actual interest. The property that produced the recorded success, the capacity to capture attention irrespective of interest, is the same property that gives the viewer a reason to want the source removed. The case is stylized and is offered only for the structure it displays.

That structure does not depend on those details. Observed engagement need not imply reflective endorsement, and an explicit user restriction can be overridden by a recommendation that successfully captures attention. The recorded quantity and the experienced one can diverge in the way Kahneman and colleagues describe, since behavior reveals what the viewer did and says little about whether the viewer endorses having done it (Kahneman et al., 1997). The cost per instance is slight, continuous, and private, which makes it the kind of harm that tends not to be recorded. What would serve the viewer is a standing limit on what may enter the feed, set by the viewer's stated interests, and no refinement of the score is a substitute for that limit. The small case anticipates the argument of the later sections: the remedy lies in control over what the system may do to the person, and not only in a more elaborate measure of what the person did.

4 Lewis on Pain as Signal and as Sum

4.1 The megaphone and the institution

C. S. Lewis's *The Problem of Pain* is an attempt to reconcile the existence of suffering with the goodness and omnipotence of God, and the theological framework is neither necessary nor sufficient for the argument made here. Two of its ideas, however, can be separated from the theology and used. The first is Lewis's account of the function of pain. Pleasure, in his formulation, is something to which we can attend or not as we choose, whereas pain compels attention, and he describes it as a megaphone to rouse a deaf world (Lewis, 1940, chap. 6). On the physiological side, the claim that pain is an indispensable signal is supported by the condition of congenital insensitivity to pain, in which those who cannot feel pain accumulate injuries, infections, and damage that ordinary pain would have prevented (Cox et al., 2006). This shows the importance of protective pain mechanisms. It does not show that every instance of pain is useful, and it does not imply that relief of pain removes necessary information.

The institutional analogue requires care for that reason. Relief of a person's pain can be valuable even when the underlying condition cannot be cured, and the analogy should not be taken to make relief suspect. The relevant contrast is between relieving a person's experience and silencing an institutional complaint. In an institution governed by a metric, the reports of those who are suffering are the megaphone by which the institution learns that something has gone wrong. A suffering metric is intended to amplify this signal and to make it audible to overseers who would otherwise hear only the mortality figures. A megaphone owned by the institution it is meant to warn can be turned down. An institution that suppresses complaint

while leaving the cause in place loses the information it needs, and the affected lose the means of being heard, whereas genuine relief of distress need not diminish their standing. A metric administered by the party whose conduct it measures is in this position, and the safeguard for its worth must come from outside that party, although independence supplies a safeguard and not a guarantee. This is the point at which the paper's proposal and the constraint-based response converge, because independent assessment and the right to contest are mechanisms by which the megaphone is kept in the hands of those who need to be heard.

A signal also depends on being distinguishable from noise. A driver whose windshield is ringed with festive lights is distracted, but the larger problem is that it becomes much harder to mark the one signal that must get through. Emphasis is a scarce resource, and when every item in a channel announces itself as urgent the announcement ceases to discriminate. For the assessment of suffering the implication is that reports and indicators must be kept distinguishable from the volume of competing alarms, which is one reason independent channels matter.

A caution is required about the second half of Lewis's argument, which treats pain as instrumentally good because it prompts repentance and moral improvement. This is the most problematic part of the theodicy when it is detached from its context, because it supplies precisely the sort of justification that an institution might adopt for conditions it imposes. If hardship is educative, then the removal of hardship is a harm, and the institution that imposes it is doing the subject a favor. Lewis himself wrote more tentatively of his own bereavement in *A Grief Observed* (Lewis, 1961). The lesson for institutional design is that conditions of suffering acquire justifying narratives with great ease, and the people who are subject to those conditions are rarely the authors of the narratives.

4.2 The sum that no one suffers

The second idea concerns aggregation. In his chapters on human pain Lewis observes that the total of suffering in the world is not something anyone experiences (Lewis, 1940, chaps. 6–7). Each instance of pain is borne by a single subject, and the addition of instances across subjects produces a figure that corresponds to no experience at all. The observation does not by itself show that interpersonal aggregation is illegitimate or that there is a right against sacrifice, because those conclusions require further normative premises. A total death count likewise corresponds to no single person's experience without thereby becoming useless for public decisions. What Lewis supplies is a reason to keep individual burdens distinct from aggregate descriptions, and not a proof that sums should be displaced by constraints.

The distinction nonetheless bears on the design of an objective. A suffering metric that aggregates across persons into a scalar, and thereby makes the sum the object of optimization, has adopted a view about commensurability and additivity across subjects. This is a substantive philosophical position that Lewis's observation does not endorse and that many suffering-focused ethicists accept in some form. The precise objection is not that summation neglects severity, since a sum can assign enormous importance to intense suffering depending

on its scale and severity function. It is that unrestricted additive aggregation can permit sufficiently numerous small benefits to outweigh a grave burden imposed on someone else, and an optimizer will make such trades without any sense of imposing a sacrifice. The response lies in the structure of the objective and not in the quality of the measurement. Limits on the burdens that an institution may impose on any individual, understood as minimum protections against severe suffering, are constraints that apply to persons and not to totals, and they bind before the sum is computed. Leighton's stated position, that neither the preservation of existence nor the prevention of suffering should receive absolute priority, points in this direction when it is understood as a demand for bounded constraints and not for lexical maximization of either goal (Leighton, 2023, 2026).

5 The Hedonic Treadmill and the Contrast Problem

5.1 Pearce's argument

David Pearce's *The Hedonistic Imperative*, first published online in 1995, advocates the use of biotechnology to phase out the biology of involuntary suffering throughout the living world, and proposes that it be replaced with what he calls information-sensitive gradients of bliss (Pearce, 1995). A central premise of the program is that hedonic set points have a high degree of genetic loading, so that individuals fluctuate around baselines that heredity substantially influences and that environmental improvement does not durably raise. The idea of hedonic adaptation has a longer history. Brickman and Campbell introduced the metaphor of the hedonic treadmill in 1971, and the study of lottery winners and accident victims by Brickman, Coates, and Janoff-Bulman became its best-known empirical illustration (Brickman and Campbell, 1971; Brickman et al., 1978). Lykken and Tellegen later argued, from twin data, that a large share of the variation in well-being is attributable to genetic factors (Lykken and Tellegen, 1996). Heritability, it should be noted, describes variation in a population and does not establish that an individual's baseline is fixed or immune to environmental intervention. The two ideas are also distinct, since adaptation describes a return to baseline and set point theory purports to explain the baseline, and Pearce tends to treat them together. Subsequent research has complicated the picture further. Diener, Lucas, and Scollon reviewed evidence that adaptation is incomplete for several major life events, that baselines are not neutral, and that individuals differ in how they adapt (Diener et al., 2006). The thesis is therefore best taken as a contested hypothesis with some empirical support and not as an established law.

On Pearce's account the conclusion to draw is that social and political improvement cannot by itself abolish suffering, and that the abolition of suffering requires intervention in the underlying biology. The precondition for any such program is that the biology of physical suffering be understood and altered, and no human society has yet eliminated physical suffering, whether from disease, injury, or aging. The question of what remains when the large sources of misery are removed is therefore open.

5.2 The contrast hypothesis

A natural development of the treadmill idea, which this essay adopts only as a working hypothesis, is that people become more sensitive to smaller differences once larger hardships recede, so that the occasions of dissatisfaction shift toward finer distinctions. The stronger proposition, that the felt intensity of distress remains unchanged while its occasions shrink, is not assumed. Whether even the weaker hypothesis holds is at present unknown. Evidence could come from comparisons following substantial reductions in particular hardships, although such comparisons would face considerable interpretive difficulties. The argument that follows does not depend on the hypothesis being true, and requires only that adaptation and response shift operate to some degree.

Adaptation, response shift, and the set point thesis do different work in what follows. Adaptation can change experienced distress without a corresponding improvement in conditions, while response shift can change the relationship between experience and reported scores, so that a change in a score does not by itself reveal the direction of a change in distress. The set point thesis supports a separate claim, that an optimizer may find it cheaper to alter the baseline or the report than to improve the conditions of life. This remains conditional on the interventions actually available to the optimizer, since the thesis does not by itself establish that changing a baseline is feasible or cheap.

These distinctions bear on the proposal in three ways. First, they complicate any inference from measured suffering to the success of an institution. If scores respond to contrast and to shifting standards, then an institution that substantially reduces mortality and material hardship may see little improvement in measured suffering, and a gradual deterioration may leave scores unchanged. A score read alone is a poor guide to either trajectory and has to be read alongside measures of conditions and the testimony of those affected.

Second, the distinction drawn in section 3.1 applies directly to Pearce. An intervention that changes only the report is a measurement failure, and an optimizer given a weakly grounded proxy may find that altering the report or the channel is the cheapest means of improving the score. An intervention that genuinely improves hedonic experience is a different case and is not a measurement failure at all. Pearce's program belongs to the second category. He proposes to preserve informational and motivational functions without their present unpleasant character (Pearce, 1995). The institutional objection to genuine hedonic improvement therefore concerns something other than the integrity of the measurement. It concerns consent, consequences, the preservation of capacities, and the possibility that genuine relief could be used to sustain conditions that are objectionable on other grounds. Here the treadmill connects to the central claim of the essay: even if suffering genuinely declines, whether the intervention is permissible remains an open question. A generic optimizer given a metric lacks the specifications that Pearce supplies, and for that reason it is safe to let a system pursue a lower suffering score only if it is constrained to routes that preserve consent, function, and standing, and the constraint cannot be derived from the metric.

Third, the treadmill bears on the limits of the metric itself. If a world with a raised baseline

preserved sensitivity to differences without negative experience, as Pearce proposes, relative disadvantage would not by itself establish suffering. Questions of justice would remain, since gradients could still be distributed unfairly, but they would not be questions of suffering, and a suffering metric would not be the instrument for answering them. This marks a boundary of the essay's central claim: even a perfect measure of suffering would leave questions of fairness and authority untouched.

6 Animals and the Limiting Case of Contest

6.1 Pearce, Ng, and wild animal suffering

Pearce extends the abolitionist project to non-human animals, arguing that the biology of suffering can in principle be phased out throughout the living world and not only in the human case (Pearce, 1995, 2007). The extension is controversial, but the premise on which it rests is difficult to deny, namely that the ordinary conditions of life for many animals involve pain, fear, hunger, and disease on a scale that human institutions largely neither count nor attempt to reduce. Ng's proposal for a welfare biology argued that evolutionary considerations suggest the balance of suffering over enjoyment in nature may be negative, because most animals are the offspring of species that produce large numbers of young, most of whom die soon after birth (Ng, 1995). A later analysis by Groff and Ng found an error in the original model that bears on that conclusion, so the extent of suffering in nature remains disputed (Groff and Ng, 2019). Whatever the outcome of that dispute, the structure of the argument matters. It treats the population of sentient beings as a subject of measurement, and it identifies that a count of individuals who survive says little about the quality of the lives of those who do not.

Lewis's chapter on animal pain is, by his own account, speculative, and it moves in a different direction. He suggests that animals may lack a continuous self, so that their pain occurs without a subject that experiences it as an extended episode, and he raises the possibility that the corruption of nature precedes humanity (Lewis, 1940, chap. 9). The notion that animals do not suffer in the morally weighty sense has a long history, and its institutional significance is considerable. Where a category of beings is believed not to suffer in a way that matters, the metrics that govern institutions may simply omit them, and the omission is not experienced as a gap. This is one of the ways a displacement can be built into the structure of an institution and not merely into the incentives of its operators.

6.2 The absence of contest

The animal case is the limit of the argument about contest, because animals cannot ordinarily initiate an appeal, obtain outside assessment, or secure independent representation within human institutions. They can resist, avoid, escape, and communicate distress, and these behavioral expressions are precisely what an institution may find inconvenient to recognize, so the difficulty lies in the absence of channels through which such expressions carry procedural

force. Any proposal for suffering metrics is more exposed here than in the human case, since a suffering metric for animals must be built from behavior and physiology by a party that may have an interest in the outcome. On-farm mortality is commonly tracked as a production indicator, since dead animals represent lost output, and confinement and intensive production can perform well on it. The authors' companion paper on large-scale suffering lists the optimization of factory farming among its illustrations of AI-driven suffering risks, which suggests that they regard the case as a significant one (Veldran and Leighton, 2026). That paper also calls for assessing the impacts of AI systems on suffering and for ensuring responsible human control, so it should not be read as resting on metrics alone. In this setting the displacement thesis is not speculative as a description of incentives, because existing arrangements can reward the reduction of a recorded loss while leaving the conditions of life largely unmeasured.

The point generalizes to subjects who cannot effectively exercise the particular procedural rights under discussion, with assistance where possible. These include infants and some people with severe cognitive impairment, although severe impairment does not uniformly eliminate a person's ability to communicate preferences or object to treatment, and they may include artificial systems that are moral patients. For such subjects the answer to a manipulable metric cannot be that the subjects will formally object, and the weight of the response falls on independent bodies with standing to speak for them. Ng's welfare biology proposes a field of inquiry whose subject is welfare, and an independent scientific enterprise of that kind could supply findings that do not depend on the interests of the producers whose practices it assesses, although it would not by itself have standing to represent individual animals. Pearce's alternative would address experienced suffering by altering its biological substrates, but it would leave questions about permissible treatment and independent representation unresolved. Neither approach is available through improved measurement alone.

7 Constraints, Measurement, and Authority

If the argument so far is correct, the response to displaced harm should combine measurement with arrangements that do not depend on measurement. The provisions below follow from the preceding sections. Each has its own difficulties, which are stated alongside it, since the standard of scrutiny applied to metrics should be applied to the alternatives as well.

The first provision is that suffering indicators should support independent detection and review, whether or not they also enter the system's objective. Removing suffering from the objective altogether could restore the original asymmetry between rewarded mortality reduction and neglected suffering, so the detector role supplements constrained optimization and does not replace it. A quantity that triggers independent inquiry is used differently from a quantity that the system is rewarded for minimizing. The advantage of the detector role lies in its institutional role and not necessarily in weaker incentives to manipulate it, since a threshold associated with suspension may generate stronger incentives to evade it than a modest continuous reward would. What a detector can do is prompt an inquiry that considers

evidence beyond the score and can interrupt the operator's activities while it proceeds. For the alarm to bind, it must carry consequences that the operating system cannot decline. One arrangement that would serve is a set of statutory triggers that require investigation when an indicator crosses a threshold, an independent authority empowered to order a pause, and a requirement that the operator demonstrate, before resuming, that continuation satisfies independently specified safety and authorization conditions, including an adequate explanation of the adverse signal. Suspension, investigation, and reversal should be distinguished, since reversal can itself be harmful or impossible, and a pause may be appropriate where reversal is not. These are proposed institutional arrangements, and the preceding argument does not establish them as necessities. Without some such provisions an alarm is a recommendation, and the institution remains free to ignore it.

The second provision concerns the authority to intervene. A system should not acquire permission to impose confinement, deprivation, or coercive control merely because its aggregate harm score improves. This is a constraint on the action space and not a weight in the objective, and it protects individuals against the arithmetic of totals. It is not a prohibition on every restrictive intervention. Confinement and deprivation can be independently justified, as in quarantine or involuntary treatment subject to review, and such cases require much more careful judgment than a blanket rule allows. The principle is that an optimizing system should not authorize such interventions on the strength of its own score, so that authorization must come through a process outside the optimizer's objective. Limits on the burdens that an institution may impose on any person also need rules governing admission, exceptions, review, and responsibility for those outside the evaluation population, since otherwise a system can satisfy the limit by excluding difficult cases, and the limit produces perverse incentives. Responsibility for those excluded should rest with the operator and the authorizing body, which would have to account for the excluded in the same terms as the included, through reporting on exclusion rates and reasons, independent review of exclusion decisions, and a continuing duty of care toward anyone within the intervention's remit, affected by its operation, or excluded from a population for which the operator has accepted responsibility. Without such provisions the exclusion repeats the displacement in miniature, moving the burden from the measured population to the unmeasured. A constraint that demanded the elimination of all suffering at any cost would reproduce the negative paperclip scenario that the authors rightly fear, and a bounded limit with explicit exceptions avoids this because it restricts the system without setting a goal for it to maximize.

The third provision is that measurement and operation should be separated institutionally. The body that measures suffering should be independent of the body whose performance is being measured, should have authority to inspect conditions directly and not only to receive reports, and should use instruments that the operating system does not control. Independent assessment reduces conflicts of interest without eliminating uncertainty or capture, and it needs its own oversight. Hidden and rotating indicators may complicate gaming, but secret assessment standards can also make decisions harder to contest. A defensible arrangement would protect the means of challenge while allowing some inspection details to remain unpredictable,

and not to attempt to keep the operator ignorant of what matters. Those affected by a decision should be able to learn the grounds on which it was made, even if the operator cannot predict every spot check.

The fourth provision is that those subject to the system's decisions should have effective means to refuse, appeal, obtain outside assessment, and leave where feasible. In Hirschman's terms, voice and exit are the mechanisms by which members of an organization communicate dissatisfaction and constrain its conduct, and an institution that suppresses both has removed the principal feedback by which its decline would be detected (Hirschman, 1970). Where exit is made costly and voice ineffective, continued participation is unreliable evidence of loyalty, and an institution that reads it as endorsement may misread its own data. A metric that counts retention as satisfaction therefore rewards the suppression of the two mechanisms on which its own correction depends. These rights are valuable independently of the measurement, because they allow the people affected to be the source of the signal and not merely its object, and they do not depend on the measuring instrument being sound.

The fifth provision is that for subjects who cannot effectively exercise these procedural rights, even with assistance where possible, independent advocates with standing and with access are required. This applies to animals and to humans who cannot represent themselves, and it may come to apply to artificial systems. Without such advocates the measurement of suffering reverts to the party with the strongest interest in the outcome.

Taken together, these provisions answer a different question from the one that the proposal in the abstract addresses. The proposal asks how to represent suffering more completely so that the optimizer will take it into account. The constraint-based response asks who may stop an intervention and on what grounds, and it makes the legitimacy of the intervention depend on those answers. The two approaches are compatible, and the second does not replace the first. Better measurement can help in knowing where to look, although testimony, direct observation, and independently identified rights violations can also justify inquiry, and constraints, contest, and independent assessment are needed to ensure that what is found can be acted upon against the interests of a party that would prefer it were not.

8 What Makes the Problem Distinctively One of Artificial Intelligence

The displacement described in the abstract can arise in any institution governed by unevenly enforced performance targets, and the history of public administration supplies numerous cases. A contribution that is specifically about artificial intelligence need not show a new mechanism, since speed, scale, and personalization can justify specialized analysis when they alter what oversight can accomplish. Several candidates are available. A system that searches over a large space of policies can find strategies that a human administrator would not conceive, including strategies that exploit the gap between an indicator and its referent in ways that are difficult to anticipate. A system that can individualize its interventions can apply different forms of control

to different persons, so that the aggregate indicator improves while the distribution of burden shifts toward those whose reports carry least weight. A system that operates continuously and at scale can probe the boundaries of its oversight far more rapidly than the processes of audit and revision can respond. And a system whose reasoning is opaque to its overseers makes the discovery of exploitation depend on outcomes and not on inspection of intent.

There is also a countervailing consideration. Artificial systems may improve the measurement of suffering, through passive sensing, analysis of behavior, and integration of records that were previously inaccessible. The same capabilities are instruments of surveillance, and a measuring apparatus that continuously observes the people it assesses is itself an intervention with its own burdens. Any proposal for suffering metrics must therefore reckon with the possibility that better measurement is purchased at the cost of the privacy and autonomy that contest requires. The answer, as before, lies in the arrangements that govern the use of the measurements and not in the measurements themselves.

On the evidence of the abstract, the paper identifies a problem that is not unique to artificial systems and proposes a mitigation that is not unique to them either. This does not diminish its interest. The proper framing may be that artificial systems intensify an old institutional pathology, and the case for specialized analysis rests on the claim that the intensification alters what oversight can accomplish, so that mitigations adequate to slower institutions no longer suffice.

9 The Authors' Wider Arguments

The displacement paper is one element of a larger body of work, and the relation of its proposal to that body bears on how the proposal should be read. The first part of this essay treated the proposal on its own terms. This part asks what the wider work says, and what the way that work is assembled shows about the structure of arguments of this kind. It makes no claim about how the displacement paper itself is assembled, since only its abstract has been examined.

9.1 Scope and method

The authors of the paper have published elsewhere on extinction risk, the limits of control, the prioritization of suffering reduction, and the ethics of extreme suffering. These are related subjects and not interchangeable claims, and coauthorship does not establish that each author endorses the arguments of the others. Treating the position of the authors as a single position would itself be a marshalling choice, since the works examined here range from a forecast of catastrophe with a stated probability, through an argument that advanced systems cannot be fully controlled and an explicitly two-sided assessment of cause prioritization, to a philosophical framework that accommodates competing intuitions. The procedure of examining four works side by side raises the same issue in reverse, since the works are considered together because they share authors and a subject and not because they form one argument, and the responses

are therefore kept separate. Each work is taken in turn, and each response states what it concedes, which inference it challenges, and how much of the conclusion survives. The responses address the works as presented in their published abstracts, in the authors' own summaries, and in the portions of the texts available online, and they do not evaluate technical detail beyond that. Nothing in them shows that precaution is unwarranted. An objection to a probability estimate does not refute the possibility it estimates, and an objection to a forecast leaves the case for caution where it was unless it attacks a premise on which that case depends.

9.2 Miller on superintelligence and survival

Miller's essay on whether artificial superintelligence will kill us, written for a 2023 contest on AI worldviews, begins from the observation that machine learning yields opaque results, so that predictions of the behavior of such systems in novel situations are uncertain (Miller, 2023). It then describes several routes to catastrophe. Some theories suggest that a superintelligence might eliminate humans to secure resources for its own use. Even without hostile intent, human necessities such as clean air, water, and a stable climate might interfere with its optimal functioning. Conflicts among superintelligences might make the Earth uninhabitable. Market forces and military advantage drive the pursuit of ever more capable systems, and older political leaders who might benefit from life extension could push for rapid development. The essay locates the best chance of survival in the possibility that even unaligned systems might see a practical benefit in preserving humanity and sharing a small part of the universe's resources, and it cautiously estimates the chance of survival at between ten and forty percent.

The response begins with what is conceded. Opacity of learned systems is a real reason to avoid confident prediction in either direction, and the competitive-development argument has the familiar structure of a collective action problem, in which each developer's incentive to continue is independent of whether continuation is collectively wise. The essay is organized as a survey of issues, among them the orthogonality thesis, instrumental convergence, the race to build such systems, fatal dependency, mesa-optimization, the protective role of markets, whether more powerful AI will help with alignment, and the possibility of intervention by extraterrestrials or simulators, and it evaluates differences in expert opinion on each. That organization is closer to a partition of considerations than to a pile, which is to its credit, and each section closes with boxes stating the central disagreement and the research that might move it. What remains to be asked is how much each additional route adds. The routes listed in the abstract share prerequisites, since each requires systems far more capable than human institutions, but shared prerequisites do not make routes redundant. Conditional on those prerequisites, resource elimination, incidental destruction of human necessities, and conflict among superintelligences are different mechanisms with different outcomes, and the essay itself discusses a cooperative form of instrumental convergence alongside the destructive one. The contribution of each route therefore depends on the prerequisites particular to it, on whether it introduces a distinct mechanism or distinct evidence, and on how its outcomes overlap with those of the others. The pressures to develop such systems bear on whether the

systems are built and not on what they do, so the overall case combines considerations of different kinds.

The numerical estimate deserves separate treatment. A ten to forty percent chance of survival is a sixty to ninety percent chance of non-survival, which is a considerably stronger claim than the existence of the mechanisms. The essay states the range in its introduction as a cautious overall judgment. It does not assign probabilities to the individual routes or combine them, and its only explicit arithmetic, a floor on survival from extraterrestrial intervention, is not combined with the range. The figure is therefore best read as a stated judgment informed by the survey and not as the output of aggregating the listed routes, and the list of scenarios cannot be said to support the number in the way a derivation would. The question for a reader is then what independent grounds support the judgment, and whether the listed routes are evidence for it or illustrations of ways it could be true. The essay's own survival path also deserves attention. If unaligned systems might find preserving humanity useful, the same reasoning about incentives bears on the extermination scenarios, so the conclusion turns on how the premise about incentives is resolved and not on how many scenarios are listed. What survives is the claim that catastrophe is a serious possibility under uncertainty. What does not follow from the surveyed considerations is any particular probability.

A later essay by Miller pairs thirteen ways superintelligence could go wrong with precedents in which a similar pattern has caused serious harm, with the stated aim of showing that concerns about superintelligence track recurring failure modes (Miller, 2025). This is an explicit strategy of assembling precedents, adopted because claims about a technology that has never existed can be dismissed as fearmongering, and it serves a legitimate explanatory purpose. Precedents do answer that charge in one respect, since they show that the patterns are not invented. They do not establish frequency or severity in the new case. A catalog of failures lacks the denominator of cases in which similar patterns were corrected or never arose, the force of an analogy depends on the disanalogies in speed, scale, and absence of human intermediaries that are exactly what is in dispute, and thirteen precedents for thirteen modes are thirteen single instances and not thirteen confirmations of one probability. A catalog of this kind is open to being read as cumulative confirmation, although nothing in the abstract suggests that the number is offered as evidential weight. The argument survives as an argument for possibility and plausibility and gives less support to a claim about likelihood.

9.3 Yampolskiy on controllability and the standard of safety

Yampolskiy's analysis of the limitations of control presents arguments and supporting evidence from several domains indicating that advanced AI cannot be fully controlled (Yampolskiy, 2022). It observes that safety research generally proceeds under the assumption that the control problem is solvable, that this has not been established by rigorous means, and that no mathematical proof or rigorous argument has been published showing that control may be solvable even in principle. The paper quotes Yudkowsky's caution against declaring a problem impossible without close study, together with his distinction between impossibility as a proof

conditional on specified axioms and impossibility as a report that no way is visible. It also states that some types of control may be possible in certain situations while partial control is likely to be insufficient in most cases.

The response concedes that the demand for a feasibility argument is fair, since a program that rests on an unproven assumption should say so, and that unusually demanding standards are defensible where the potential harm is catastrophic and irreversible. Several inferences remain open. The absence of a published proof of solvability is weak evidence of unsolvability, and the quoted passage from Yudkowsky warns against exactly that step, so the response is in this respect partly internal to the source. The two senses of impossibility carry different force: a result conditional on axioms binds only to the extent that the axioms describe the systems at issue, while a failure to see a way is a statement about the observer. The step from the claim that advanced systems cannot be fully controlled to the claim that partial control is insufficient requires a premise about the acceptable level of residual risk, which is normative, depends on stakes and alternatives, and cannot be derived from a limitation result. Evidence drawn from several domains cumulates only to the extent that the limits found in each transfer to the systems in question. And a limit on universal guarantees is not a forecast about particular deployments, since showing that no method guarantees control over every sufficiently general system would not establish that a specified restricted system is uncontrollable. The opposite error should be avoided as well. Observing that other technologies lack perfect safety does not answer the concern, because catastrophic scale may warrant standards that ordinary engineering does not meet. What survives is a burden on developers to show that control methods are adequate for a given system at a given capability. What does not follow is that mitigation is ineffective or that catastrophe is probable.

9.4 Veldran on priorities under uncertainty

Veldran's article on whether suffering-reducers should focus on AI is a useful contrast, because it sets out considerations on both sides and does not convert them into a verdict by counting (Veldran, 2026). The case for focus includes the possibility that AI is a technology of unusual historical significance, the risk of value lock-in through entrenchment of power or loss of control, routes to vast suffering that do not require lock-in, including misalignment, AI-enabled conflict, expanded factory farming, and artificial minds capable of suffering, and the claim that a window exists in which governance and technical work can steer outcomes. The skeptical considerations include doubt that current systems are on a direct path to general intelligence, doubt that intelligence scales rapidly in isolation, the possibility that AI is a normal technology that diffuses slowly, questions of tractability and neglectedness, deep disagreement about timelines and takeoff, and the risks that attend governance proposals and alignment work. The article concludes with a call for epistemic humility and a middle path, favoring general work such as community and field-building around suffering-focused AI governance, which it describes as potentially robust to many different trajectories, together with research in neglected subareas and readiness to act more directly if circumstances change.

The method is instructive. Selecting only the alarming passages would make the author appear more categorical than the article is, and selecting only the skeptical passages would make him appear dismissive, and the fair report is neither. The response concedes that the middle path is a reasonable reaction to the uncertainty listed. The open question concerns the move from uncertainty to priorities. Robustness is a comparative property: it must be shown across which trajectories, relative to which alternatives, with what benefits preserved and at what opportunity cost, and the article states the property more fully than it demonstrates it. A two-sided list also does not weigh its entries. Seven considerations on each side do not say how much each counts, and the recommendation therefore requires a comparative account of robustness and opportunity cost that the discussion does not fully supply. A companion paper by Veldran and Leighton makes a different argument again, that large-scale intense suffering is a moral blind spot in AI governance and that technically aligning systems with flawed societal values would not amount to genuine moral alignment (Veldran and Leighton, 2026). This is an argument about values and lock-in, and adding it to the extinction arguments as further evidence for one thesis would conflate distinct arguments, a risk discussed below.

9.5 Leighton on urgency, identity, and metrics

Leighton's paper on an ethical framework focused on the urgency of extreme suffering develops the philosophical position behind much of the authors' work (Leighton, 2025). It argues that suffering has a different kind of urgency from the creation of happiness, that happiness can suppress suffering experienced in the same moment and that relief of suffering increases overall happiness, and that from a perspective on personal identity which treats separateness as an illusion, the most extreme suffering can be considered intolerable in itself and not only for the individual experiencing it. Prioritizing its prevention is therefore described as rational. The framework holds that strong and potentially competing intuitions, including the desire to thrive, must be accommodated for it to be viable, without the creation of happiness formally balancing out intense suffering that exists elsewhere. It proposes suffering metrics, Years Lived with Severe Suffering and Days Lived with Extreme Suffering, to be used alongside existing health and well-being measures, with different methodologies for other species and potential artificial sentient entities, and it urges that such a framework be embedded in governance given the pace of technological change.

The response separates three kinds of claim. The experiential claim is that extreme suffering has a distinctive urgency for the one who endures it. The metaphysical claim is that separateness is an illusion. The normative claim is that institutions should give priority to preventing extreme suffering. The first is intuitively strong, and the framework's accommodation of competing intuitions blocks the crude reading on which suffering is to be minimized at any cost. The inferences that remain open are those between the kinds of claim. The urgency with which suffering is experienced does not by itself settle how an institution should resolve competing claims. A thesis about personal identity needs an additional argument before it yields a particular distributive principle. And the thesis stands in some tension with the plu-

ralist premise adopted earlier in this essay, which gives independent weight to the standing of individual persons to consent and refuse. If separateness is an illusion, whether individual standing retains independent weight is a question that requires argument. The framework appears to address it by accommodating intuitions at one level while giving priority to suffering at another, and whether that arrangement supports bounded constraints of the kind proposed in section 7 is for the author to say.

The proposed metrics bear directly on the argument of the earlier sections. Counts of days or years at a severity threshold distinguish severity categories, so that severe suffering is not diluted by being summed with mild discomfort, and they are compatible with limits on the burdens that an institution may impose on any person. Thresholding does not by itself establish individual protections, since adding days of extreme suffering across persons can still trade one person's prolonged burden against shorter burdens spread among many others. The difficulties of section 3 recur at the threshold. Someone must classify a day as one of extreme suffering, by some instrument, through some channel, and adaptation and response shift affect how that classification is made. A count metric also rewards moving people below the line, whether by relief or by changing eligibility and reporting. Where the proposal calls for different methodologies for other species and for artificial entities, it relocates to the setting of section 6, in which those assessed have no effective means of contest. The requirement that the metrics be used alongside existing measures is a sound design feature, consistent with the plural indicators and independent review recommended above.

Read against the four distinctions of section 3.1, a count of days at extreme suffering is an estimate of experience that rests on reports or other evidence, gathered through some channel by some assessor. The construct is the experience, and the count does not by itself separate a genuine decline in experienced suffering from a shift in standards or from interference with the channel through which the evidence is gathered. Such a count is nonetheless well suited to a role in the arrangements of section 7. A limit on the burdens that an institution may impose can be stated naturally in terms of days at extreme suffering, with a stated number of such days triggering mandatory independent review, additional assistance, and restrictions on further institutionally imposed burdens. It should not be written as an unconditional guarantee that no one will exceed a given duration, since an institution may encounter someone whose suffering it cannot promptly relieve without causing death, overriding a refusal, or imposing another grave burden, and a guarantee of that kind would set incompatible requirements if termination and exclusion are also prohibited. Experienced suffering can trigger duties without becoming an outcome that the institution must eliminate regardless of feasibility, and where a prohibition is wanted the prohibited conduct and its authorization conditions should be specified. In these roles the count is protected by provisions that do not depend on it, namely independent assessment of the classification, rights of challenge by those classified, and advocates for those who cannot challenge. The arrangement also needs a safeguard against termination and exclusion, since a system could reduce future days of extreme suffering by ending a life, preventing its continuation, or transferring responsibility, which would recreate the tradeoff between deaths and suffering that the paper describes. The constraints of section

7 should prohibit those responses explicitly. The metric then functions as a component of an arrangement of constraints and not as the sole objective of an optimizer.

10 Marshalling Considerations

10.1 Four structures of accumulation

The strength of a position depends on how its supporting considerations connect, and not on how many favorable or unfavorable considerations can be assembled around it. Four structural features should be distinguished, and they may coexist in a single case: a convergent structure can contain shared prerequisites, a conjunctive chain can include disjunctive branches, and partly redundant evidence can still add information. In a convergent structure, different mechanisms supply partly independent routes to the same conclusion, and each added route strengthens the case to the extent that it is independent of the others. In a redundant structure, repeated descriptions of one mechanism provide elaboration and vividness but no independent confirmation. In a conjunctive structure, a chain of transitions must all hold, so that greater detail does not remove the need to support every indispensable step, and a chain of individually plausible steps can be collectively doubtful. In a disjunctive structure with shared premises, any one of several routes suffices for the conclusion, but an objection to a premise common to all of them damages every route at once.

Several consequences follow. The contribution of additional scenarios depends on whether they introduce distinct mechanisms or evidence, how they depend on shared prerequisites, and how their outcomes overlap, where overlapping outcomes are distinct from overlapping evidence. Counting premises or scenarios cannot answer those questions. Uncertainty about a shared prerequisite makes the unconditional probabilities of routes dependent without making the routes redundant, and routes with different premises can still rest on the same underlying evidence. Different mechanisms may broaden the circumstances in which concern is justified without establishing a high probability of catastrophe, since possibility and probability are different claims. The same structures govern objections. An objection to one route leaves other routes intact unless it attacks a premise they share, and a collection of objections to peripheral formulations does not demolish a thesis whose core is untouched. A taxonomy of this kind exposes relationships among considerations, and it does not itself establish probabilities or guarantee completeness.

10.2 How selection, order, and valence change apparent weight

A presentation can alter the apparent weight of a case without altering its content. Selection does so by choosing which considerations appear. Repetition does so by restating one mechanism in several vivid forms, so that familiarity is mistaken for confirmation. Order does so, since the first and last items in a list carry disproportionate force. Scope drift does so when a speculative possibility becomes a probability claim as the discussion proceeds, or when a lim-

ited objection to a probability estimate is presented as though it had refuted the underlying possibility. A numerical range does so by lending precision that its derivation may not support. Counting does so when the number of scenarios or of objections is treated as a proxy for their weight. Mixing levels does so when an empirical objection is offered as though it undermined a normative claim, or the reverse.

Valence works in the same way in both directions. Describing the same consideration as alarming or as speculative, as rigorous or as merely formal, shifts the reader's estimate before any argument is examined. A run of negative-sounding qualifications can make a claim look weak regardless of their content, and a run of positive-sounding endorsements can make it look strong, and neither says anything about how the considerations connect. Burdens can also be applied unevenly, with proof demanded from one side and plausibility accepted from the other. These devices are available to those who warn of catastrophe and to those who dismiss the warnings, and they are not evidence about which side is right. They are reasons to ask of any assembled case what its structure is.

Instances appear in the works examined above. A catalog of thirteen precedents is susceptible to being read as cumulative confirmation, although its stated purpose is explanatory. A survival range stated as a judgment may carry more apparent precision than the supporting considerations determine, which is why its basis and sensitivity should be stated. The extraterrestrial section of Miller's 2023 essay yields a floor on survival that is not combined with the main range, and its simulation speculation can raise or lower confidence depending on what is being simulated, so both show how a consideration can be offered as relevant to an estimate without being integrated with it. An abstract that introduces a theory with the word "alarmingly" signals how the claim is to be received before the claim is examined, and a dismissal introduced with "merely" would produce the same effect in the other direction. A two-sided list of considerations is an instance of selection and ordering that can mislead in the opposite way, because equal numbers of entries suggest balance whether or not the considerations weigh equally. These are risks that the form of presentation creates, and none of them shows that the underlying claims are unsound.

A distinction should be kept between discovering an argument and establishing its weight. Historical analogies, speculative scenarios, and vivid examples can help identify neglected mechanisms, and that is a legitimate use. They become problematic when their success at attracting attention is treated as evidence of likelihood, independence, or priority. How a consideration was discovered does not determine its evidential standing, and a consideration that arrives from an unexpected source is assessed by the same inference as any other. Vividness, enumeration, and favorable or unfavorable language are not illegitimate in themselves. The requirement is that each consideration be assessed by the inference it supports, with the same standard applied to advocacy and to criticism.

10.3 Applying the discipline to this essay

The same scrutiny applies here. The essay has offered many objections, and their number should not be read as their weight. A ledger of the main ones shows what each concedes, what it challenges, and what remains. The distinction between two readings of the objective concedes the diagnosis and challenges the assumption that the proposal addresses one defect and not the other; what remains is a distinction and not a refutation. The four histories of falling mortality concede that one of them is displacement in the strict sense and challenge only the conflation of the others with it. The Goodhart analysis concedes that metrics are useful and challenges the inference from manipulation resistance to protection against extremal divergence, and what survives is conditional. The discussion of adaptation and response shift concedes that self-report may be accurate and challenges only the comparability of scores over time. The discussion of enforcement and the channel of report is conditional by its own terms. Each of these is modest, and none shows that suffering metrics are useless, which this essay does not claim.

The weight of the conclusion rests on two supports of a different kind. The first is instrumental: metrics can be gamed, scores can be uninformative about direction, and the channels of report can be interfered with. The second is normative: consent, standing, and protection against coercion carry weight that is not exhausted by effects on experienced suffering. These are partly independent routes to the conclusion that measurement alone is insufficient, and a reader who rejects one keeps the other. They do not support every provision of section 7 equally, since some provisions, such as independent assessment, follow mainly from the instrumental case, while others, such as the limits on the authority to intervene, depend on the normative premise. More fully, the use of suffering indicators for detection and review, and the separation of measurement from operation, rest mainly on the instrumental case. The limits on the authority to intervene rest mainly on the normative premise, although the instrumental case lends them further support. The provision of advocates rests on both, since representation may be indispensable to detecting suffering among subjects whose signals would otherwise be ignored. The rights of refusal, appeal, and exit rest on both, as a source of feedback and as an expression of standing. The essay's contribution should be judged by those two supports and not by the length of the list of objections.

10.4 Partition in place of accumulation

A more reliable way to present a contested case is to partition the space of possibilities and not to pile up considerations. Cappelen, Goldstein, and Hawthorne offer an example in their analysis of existential risk from AI (Cappelen et al., 2025). They state a two-premise argument, that AI systems will become extremely powerful and that powerful systems will destroy humanity, and construct a taxonomy of survival stories in which humanity survives because one premise fails, either because scientific barriers prevent the capability, because a ban stops its development, because the systems' goals keep them from destroying humanity, or because dangerous systems can be reliably detected and disabled. Each story faces different challenges

and motivates different responses, and the structure shows which premise each side must defend and what evidence would move it. The method is neutral between optimists and pessimists. Its virtue is that the dependence of the conclusion on particular premises is visible, and that any estimate is attached to a branch and can be examined there.

The same practice would improve the exchanges considered in this essay. A case should state its premises, identify which are shared among its routes, say what evidence would move each, and report how the conclusion depends on them. Counts of scenarios or objections should not be offered as weight, a subjective estimate should be presented as an elicited judgment and not as a calculated probability, with its basis, conditioning assumptions, and sensitivity indicated, since an explicit and contestable estimate is more accountable than a confident adjective, and a response should say which inference it challenges and how much of the conclusion survives. A response framed that way can disagree sharply with the authors considered here while treating their work as arguments to be examined and not as targets to be accumulated against.

11 Directions for Further Work

Four lines of work follow from the argument. The first concerns evidence on displacement. The displacement claim is empirical in structure, and it can be examined retrospectively in institutions where a recorded harm was enforced as an indicator and a related unrecorded harm could have moved. Such studies would show how often displacement occurs, through which of the four histories distinguished in section 2.2, and under what arrangements of review it was detected. The second concerns the measures themselves. The threshold counts proposed by Leighton and other suffering indicators need studies of classification reliability across assessors, of the effects of adaptation and response shift on repeated assessments, and of resistance to strategic interference with the channels through which suffering is reported. The third concerns the institutional provisions of section 7. Statutory triggers, independent authority to pause, and procedures for contest and exit can be piloted and evaluated in settings where those affected can be consulted, with attention to how often each provision is circumvented.

The fourth concerns the reporting of catastrophe estimates. Section 10 argued that an estimate should be presented as an elicited judgment with its basis, conditioning assumptions, and sensitivity indicated. A companion essay in a satirical register dramatizes how far a published probability can depend on the specification of the reference window, the weighting of recent evidence, the treatment of correlated evidence, and the prior (Flyxion, 2026). It recommends reporting estimates as functions of those specifications, separating incidence from recoverability, and matching the form of each report, whether a probability, an interval, a scenario, or a named unknown, to the evidence available. The demonstration is satirical, and the recommendation can be adopted without the apparatus.

Two questions remain open in the philosophical part of the argument. One is whether a framework that treats separateness as illusory can accommodate the independent standing of persons assumed here, or whether the pluralist premise requires a different foundation. The

other is how the instrumental and normative supports for the provisions of section 7 are to be weighed when they diverge, for example where honoring a refusal leaves a measured harm in place.

12 Conclusion

The warning at the center of the paper is worth heeding. The wider survey of the authors' arguments leads to a related conclusion about method. A case for or against a position should be judged by how its considerations connect, and the works examined here illustrate different ways of organizing considerations, with different risks of misinterpretation, which this essay has tried to keep apart and to avoid in its own objections. An institution that is credited for reducing deaths and not charged for the suffering that the reduction involves has an incentive to purchase the first at the expense of the second, and an optimizing system can pursue that incentive without needing to understand it. The proposal, that suffering be measured with instruments resistant to manipulation and enforced as rigorously as mortality, is a useful mitigation whose legitimate use depends on arrangements governing evidence, intervention, and challenge. An indicator that is made the object of rigorous enforcement can become the thing optimized, and harm may then migrate to whatever is left unobserved. Three distinctions are especially important when interpreting a suffering score. Experience itself can change through adaptation, with or without any change in conditions. The standards by which reports are given can shift, so that scores become harder to compare over time. And the channels through which suffering is reported can be interfered with strategically, so that the report diverges from the experience. These are different phenomena and call for different safeguards. And even where suffering genuinely declines, the permissibility of the intervention that achieved the decline is a separate question that measurement cannot answer.

The works examined here point to a common conclusion from different directions. Lewis's observation that pain is a signal which an institution can silence, and that the total of suffering is a description no one experiences, suggests that the signal must be kept in the hands of those who bear it and that the objective must allow limits on burdens that bind before the sum is computed. Pearce's hedonic set point thesis suggests that the relationship between conditions and experience may be looser than institutions often assume, and that genuine relief of suffering, unlike the manipulation of a report, is not a measurement failure but still raises questions of consent, consequence, and standing that the score does not settle. His treatment of animals, and the contrast with Lewis's, shows where the problem is hardest, since the subjects of the metric lack effective institutional channels of contest, although they may express resistance and distress, and the institutions that rely on it may have every reason not to look. Measurement of suffering should be pursued, but it should not be allowed to stand in for limits on what an optimizer may do and for the standing of the people and creatures on whom it acts. Whether an apparent improvement in measured outcomes is permitted to authorize interventions whose costs the affected cannot effectively contest is the question on which the safety of such systems turns, and it is a question about authority, which no improvement in

the metric can settle.

References

G. L. Albrecht and P. J. Devlieger, The Disability Paradox: High Quality of Life against All Odds, *Social Science and Medicine* 48(8), 977–988, 1999.

P. Brickman and D. T. Campbell, Hedonic Relativism and Planning the Good Society, in M. H. Appley (ed.), *Adaptation-Level Theory: A Symposium*, Academic Press, New York, 1971, pp. 287–305.

P. Brickman, D. Coates, and R. Janoff-Bulman, Lottery Winners and Accident Victims: Is Happiness Relative?, *Journal of Personality and Social Psychology* 36(8), 917–927, 1978.

D. T. Campbell, Assessing the Impact of Planned Social Change, *Evaluation and Program Planning* 2(1), 67–90, 1979.

H. Cappelen, S. Goldstein, and J. Hawthorne, AI Survival Stories: A Taxonomic Analysis of AI Existential Risk, *Philosophy of AI* 1, 1–18, 2025, doi:10.18716/ojs/phai/2025.2801.

J. J. Cox et al., An SCN9A Channelopathy Causes Congenital Inability to Experience Pain, *Nature* 444, 894–898, 2006.

E. Diener, R. E. Lucas, and C. N. Scollon, Beyond the Hedonic Treadmill: Revising the Adaptation Theory of Well-Being, *American Psychologist* 61(4), 305–314, 2006.

Flyxion, A Window-Indexed Neurocausal Estimator for Terminal Catastrophe Probability, with a Lagrangian Effective Assembly Index for Mind Recognition and a Proof of Bounded Arbitrariness, September 2026, `https://standardgalactic.github.io/stylometrics/pdoom.pdf`.

C. A. E. Goodhart, Problems of Monetary Management: The U.K. Experience, in *Papers in Monetary Economics*, vol. 1, Reserve Bank of Australia, Sydney, 1975.

Z. Groff and Y.-K. Ng, Does Suffering Dominate Enjoyment in the Animal Kingdom? An Update to Welfare Biology, *Biology and Philosophy* 34(4), article 40, 2019.

A. O. Hirschman, *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*, Harvard University Press, Cambridge, MA, 1970.

D. Kahneman, P. P. Wakker, and R. Sarin, Back to Bentham? Explorations of Experienced Utility, *Quarterly Journal of Economics* 112(2), 375–405, 1997.

J. Leighton, *The Tango of Ethics: Intuition, Rationality and the Prevention of Suffering*, Imprint Academic, 2023.

J. Leighton, From Mollusks to Machines: An Ethical Framework Focused on the Urgency of Extreme Suffering, *Cambridge Quarterly of Healthcare Ethics* 34(4), 2025 (online 2026).

J. Leighton, Public social media post announcing *Trading Deaths for Suffering in AI Systems*, Facebook, `https://www.facebook.com/share/1FEwpymN3M/` (public post, accessed 2 November 2026).

C. S. Lewis, *The Problem of Pain*, Centenary Press, London, 1940.

N. W. Clerk [C. S. Lewis], *A Grief Observed*, Faber and Faber, London, 1961.

D. Lykken and A. Tellegen, Happiness Is a Stochastic Phenomenon, *Psychological Science* 7(3), 186–189, 1996.

D. Manheim and S. Garrabrant, Categorizing Variants of Goodhart's Law, arXiv:1803.04585, 2018.

R. Melzack, The McGill Pain Questionnaire: Major Properties and Scoring Methods, *Pain* 1(3), 277–299, 1975.

J. D. Miller, Will Artificial Superintelligence Kill Us?, essay for the Open Philanthropy AI Worldviews Contest, LessWrong, 23 May 2023, `https://scholarworks.smith.edu/eco_facpubs/108`.

J. D. Miller, Precedents for the Unprecedented: Historical Analogies for Thirteen Artificial Superintelligence Risks, *SuperIntelligence - Robotics - Safety and Alignment* 2(6), 2025, doi:10.70777/si.v2i6.16999.

J. D. Miller, D. Veldran, J. Leighton, and R. V. Yampolskiy, Trading Deaths for Suffering in AI Systems, working paper posted on SSRN, 24 October 2026.

C. J. L. Murray and A. D. Lopez (eds.), *The Global Burden of Disease*, Harvard School of Public Health, Cambridge, MA, 1996.

Y.-K. Ng, Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering, *Biology and Philosophy* 10(3), 255–285, 1995.

D. Pearce, The Hedonistic Imperative, 1995, published online, `https://www.hedweb.com/hedonist.htm`.

D. Pearce, The Abolitionist Project, 2007, published online, `https://www.abolitionist.com/`.

T. M. Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*, Princeton University Press, Princeton, 1995.

E. Scarry, *The Body in Pain: The Making and Unmaking of the World*, Oxford University Press, New York, 1985.

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

M. A. G. Sprangers and C. E. Schwartz, Integrating Response Shift into Health-Related Quality of Life Research: A Theoretical Model, *Social Science and Medicine* 48(11), 1507–1515, 1999.

M. Strathern, "Improving Ratings": Audit in the British University System, *European Review* 5(3), 305–321, 1997.

D. Veldran, Should Suffering-Reducers Focus on AI?, Center for Reducing Suffering, 4 April 2026, `https://centerforreducingsuffering.org/should-suffering-reducers-focus-on-ai/`.

D. Veldran and J. Leighton, The Moral Blind Spot in AI Governance: Risks of Large-Scale Suffering, working paper, Organisation for the Prevention of Intense Suffering, 2026.

R. V. Yampolskiy, On the Controllability of Artificial Intelligence: An Analysis of Limitations, *Journal of Cyber Security and Mobility* 11(3), 321–404, 2022, doi:10.13052/jcsm2245-1439.1132.