

We Measure the Evidence. You Make the Call. Here's Why

Wladimir J Alonso & Cynthia Schuck-Paim

Recommended citation

Alonso, W. J. & Schuck-Paim, C. (2026). WelfareWe Measure the Evidence. You Make the Call. Here's Why. Welfare Footprint Institute Position Paper Series. 261

We Measure the Evidence. You Make the Call. Here's Why.

March, 2026

Table of Contents

Executive Summary	2
Table of Contents	3
Introduction	4
Two kinds of criticism, and why we treat them differently	4
1. Methodological and evidentiary criticism	4
2. Valued based objections	5
Example: cage and cage-free comparison: the need for quantification	6

Introduction

We often receive feedback from people who share a common goal: improving animal welfare as much as possible. When feedback is in the form of critique, it may come in two forms. One kind is scientific: about evidence, modeling choices, uncertainty, and gaps. That sits squarely within our scientific mandate, and we treat it as valuable input for revision. The other is normative: about what conclusions people should draw, or what actions people should take, given the evidence. That's a legitimate debate, but it isn't a scientific dispute, and it isn't where an evidence-focused research institute is best placed to contribute.

This post is a standing clarification of principles: how we interpret criticism, what we do with it, and what we see as our role in the broader effort to improve animal welfare.

Two kinds of criticism, and why we treat them differently

In practice, critiques often get mixed together. Separating them makes the conversation more productive.

1. Methodological and evidentiary criticism

Our lane: building evidence-based, auditable welfare metrics

The Welfare Footprint Institute exists to do a specific kind of work: translate the best available scientific evidence about animals' lived experiences into transparent, comparable welfare metrics, grounded in explicit assumptions, accompanied by uncertainty where appropriate, and open to revision as more evidence emerges.

We don't aim to deliver a "final answer" on welfare impacts. We aim to produce auditable, evidence-based metrics that remain explicitly conditional on their assumptions, and therefore improvable over time. This approach is aligned with Effective Altruism's core principles: impartiality, careful reasoning under uncertainty, epistemic humility, and improving decisions to reduce suffering as effectively as possible.

Methodological critiques may include claims that a given module of the WF framework (e.g., the description of living circumstances) is not capturing reality well, or that a specific analysis used questionable assumptions, drew boundaries too narrowly, or misread the evidence on intensity, duration, or prevalence. This is exactly the kind of criticism we want. Where warranted, we address it the way science is supposed to work: by revising assumptions, expanding inventories, refining estimates, and/or widening

uncertainty ranges. That is not a threat to the framework; it is precisely how the framework is meant to function. In that sense, we aim for what Nassim Nicholas Taleb calls “antifragility”: a system that doesn’t merely tolerate stress-testing, but improves because weaknesses can be exposed and corrected.

A note on scope, completeness and validity

Every assessment defines a specific scope of analysis depending on the aim. That means critiques may accurately point out that an analysis is not exhaustive. But “not exhaustive” does not mean “invalid.” A proper claim of invalidity requires more than identifying what was left out — it requires showing that the omitted components (harms, life fates, life phases) lead to sufficiently prevalent, intense, and long-lasting experiences that can plausibly reverse the direction of claims or materially change the magnitude of the conclusion. Without that, the critique is an observation about scope, not a legitimate refutation of the result. We include an example of how this works in practice at the end of this post.

2. Valued based objections

A different type of feedback is concerned not with whether an estimate is accurate, but with how results might be interpreted socially or politically. Given the same welfare estimate, reasonable people can arrive at different ethical conclusions:

- Some may focus on comparative harm reduction: “This option improves welfare substantially; it should be prioritized.”
- Others may hold threshold-, rights-based, harm prevention views: “Any non-trivial amount of severe suffering is unacceptable; abstinence is the only acceptable choice.”
- Still others may weigh trade-offs differently depending on uncertainty, moral risk tolerance, or strategic constraints: “Animal welfare needs to be viewed within the broader decision-making context”.

We understand why people can have these concerns: Those prioritizing harm prevention can worry that quantified and demonstrable improvements in animal production systems could be used to defend continued consumption, while others may worry that quantified harms will be used to justify stronger demands being placed on producers, objections to animal production practices or systems, or increased product costs. These concerns often have sincere ethical or strategic origins, but they are not, at root, a scientific dispute about the correctness of a welfare estimate. They are disagreements about ethical thresholds, strategy, and messaging.

Like everyone else, the people involved in the Welfare Footprint Institute have personal values and moral intuitions, but as an Institute, we do not take an institutional position on which ethical framework is “correct.” Quantifying the welfare impact of a system is not a declaration that the system is morally acceptable. It is an attempt to describe, as accurately as possible, what the animals are likely experiencing. In fact, one reason to build time-based welfare metrics is that they can make important welfare improvements possible, by translating them into units that are visible, concrete, comparable, and auditable.

As an Institute, we deliberately avoid presenting welfare measurement as a moral verdict. The central hope is simple: that better measurement helps make animals’ lives better by improving prioritization, revealing hidden hotspots of welfare loss, clarifying trade-offs, and making welfare impacts transparent and relatable to a wide set of decision-makers. Our role is to make the welfare-relevant evidence base clearer, so that individuals and organizations, whatever their ethical lens, can reason from a shared empirical foundation.

Example: cage and cage-free comparison: the need for quantification

In 2021, we analyzed the welfare [impact](#) of transitioning from cages to indoor cage-free aviaries. For each housing system, the analysis thus focused on the laying phase, and examined key welfare challenges likely to be affected by the transition. Challenges that are equally shared across systems, such as hatchery processes, would not change comparative results and overall conclusions and were therefore not included in the analysis.

The comparison quantified cumulative time spent in negative affective states and incorporated major welfare harms known to differ across systems, including chronic behavioral deprivation in cages and higher prevalence of certain injuries in cage-free systems. The key finding was that cage-free aviaries were superior to cages in terms of hours of pain prevented even soon after a transition to cage-free environments. As with any welfare assessment, the scope was explicit and bounded: to estimate the welfare impact of the cage-free transition, not the impact of a whole life as an egg-laying hen in a system.

Importantly, the assumptions and scope of this first analysis favored a more conservative estimate of the welfare benefits of cage-free housing. First, while we did not include harms associated with poor litter or air quality in poorly managed aviaries, we also did not consider several welfare costs of cages: increased levels of fearfulness, the absence of agency and control (conducive to learned helplessness), induced molting (still applied to caged layers in some countries but rarer in cage-free production), longer

production cycles associated with end-of-lay deterioration, and the rearing phase, which if included would amplify estimated differences. Second, the analysis assumed that a given injury or disease was associated with the same duration and perceived intensity of pain regardless of housing system. [Substantial evidence](#) indicates otherwise, suggesting that similar welfare challenges produce more intense and longer-lasting pain in cages. Third, the prevalence data available for cage-free systems reflected facilities still in transition, before management experience had caught up with the decades of optimization behind caged production, so higher prevalence values of harms were assigned to cage-free systems. The welfare benefits of the cage-free transition are therefore likely to surpass the estimates from this first analysis. Ongoing work with a different goal — estimating the full welfare footprint of an egg — will extend these estimates to 100+ affective experiences across the full production cycle (breeding farms, hatcheries, production farms and abattoirs).

One of the motivations for this work is the widespread assertion that no housing system is objectively better because each has its own welfare challenges. This assertion confuses the existence of trade-offs with their equivalence: the fact that each system has problems does not mean the impact of the problems are equal. Two systems can each have serious welfare challenges and still differ enormously in their overall animal welfare impact, because harms differ in how painful they are, how long they last, and how many animals they affect. Without quantifying these dimensions, any comparison remains an exercise in list-making. For example a chronic deprivation affecting every animal in a flock can be made to look equivalent to an injury affecting a small fraction of animals for a short time. The framework exists precisely to resolve this — by quantifying the intensity, duration, and prevalence of each harm so that comparisons rest on evidence rather than assertion.