cite as "Treves, A. Standards of evidence in wild animal research. Report for the Brooks Institute for Animal Rights Policy & Law. First published 30 June 2019, modified 4 January 2020
http://faculty.nelson.wisc.edu/treves/CCC.php

Standards of evidence in wild animal research

Table of Contents

Abstract

This document is designed as a list of principles and expectations for gold standard research on wild animals. It is intended for those funders, scientists, peer reviewers, editors, publishers, or reporters who are supporting, conducting, reviewing, or communicating research to any audience. Stated simply, gold standard research aims for the strongest inference conducted with the highest standards of evidence and scientific integrity.

First, research should adhere to 4 essential principles of scientific integrity.

- Transparency is the clear and thorough explanation of all assumptions, methods, and steps in science. As a precondition and consistent pattern, every step in the research process should be clear and understandable to an educated lay audience. The principles of objectivity, reproducibility, and independent review depend on thoroughgoing clarity. Therefore, each of the following steps should pass its own test of transparency.

- Objectivity is "the ability to consider or represent facts, information, etc., without being influenced by personal feelings or opinions; impartiality; detachment" [1]. Starting assumptions, worldviews, and presuppositions should be made explicit at the outset, beginning with anthropocentric or non-anthropocentric value judgments (does the researcher grant humans and nonhumans equitable consideration or place priority on humans or nonhumans?) In addition, the researcher should make clear if legal structures or institutional permits relating to property rights or responsibilities toward animals have shaped their research design and be explicit about which legal requirements act under what jurisdiction. Objective research almost always includes statement of opposed alternatives, as in "We tested x against its alternative(s) y and z" rather than "We tested the effectiveness of x". Also, y and z should be genuine plausible alternative explanations. Even for research that is not experimental, it is wise for scientists to keep in mind alternative explanations for cause-and-effect or for the origins of natural phenomena throughout the research process. Note that some research on animals involves animals as interventions in addition to animals as subjects (e.g., predator-prey experiments). In such experiments, all of the recommendations in this document should be considered for both the treatment animals and the subject animals. For simplicity, we refer to subjects below for all animals involved in research.

- Reproducibility is the quality of a scientific finding that can be replicated by other scientists under the same conditions as the original: Independent scientists should be able to follow written descriptions of research methods to replicate every step and the findings, given sufficient resources and equipment. Facilitating reproducibility is a responsibility of the original researcher and they should welcome oversight of that facilitation and welcome efforts at replication, including sharing materials, techniques, and raw data no matter how intellectual property is conceived and despite rivalry or interpersonal animosities. Failure to reproduce is a bad sign for the evidentiary strength of the original research if the effort at replication is done in good faith with care. There are three categories of reproducibility: exact, technical, and conceptual. Exact reproducibility requires every step in the original process be replicated identically, which is rare because the location, timing, materials, individuals, etc. might be influential on the findings and might differ in the subsequent replication efforts. Failure to replicate under exact reproducible research suggests the original findings were misleading. Technical reproducibility is more common and consists of replicating with very close approximations of all methods. Failure to replicate under such circumstances might

[1] (OED, 2018)

require review of the described methods and repetition by one or both parties. Finally, conceptual reproducibility aims to replicate the findings by a different cause-and-effect pathway or using different methods. Such efforts can be powerfully confirmatory of underlying biological mechanisms exposed by the original research. Failure to replicate might indicate the causal mechanism was misidentified by the original researchers or the subsequent researchers erred.

- Independent review: Before data collection, scientists should subject the proposed methods to scrutiny and subject their own interpretation of data after their collection to scrutiny. Publicizing scientific communications prior to independent review is a questionable practice, although this is an evolving debate in the literature on pre-publication review. The scrutiny of methods and scrutiny of interpretations should be undertaken by qualified parties with an arm's length relationship to the researcher and without conflicts of interest about the scientists conducting the research or their findings. Conflicts of interest relate mainly to financial or career advancement issues, not to differences of opinion. Researchers should welcome review by experts in their field, not side-step such review by omitting citation to such experts or by explicitly discouraging those experts as reviewers. Peer reviewers should also follow the steps in this document, particularly for transparency and objectivity. Reviewers, editors, and publishers also have specific responsibilities for the quality of scientific communications. The specific responsibilities relate to maintaining the quality of the scientific record long after a particular scientific communication has been made public and passed review (See the 2019 guidelines of the Committee on Publication Ethics (COPE https://publicationethics.org/about/our-organisation accessed 1 June 2019.). Researchers have primary responsibility for correcting, retracting, or publicly expressing diminished confidence in their own scientific communications no matter how old they might be. The broader scientific community has secondary responsibility for cleaning the scientific record if credible evidence surfaces of omissions, errors, or misconduct (fabrication, falsification, or plagiarism). Efforts by any party to silence critics or ignore qualified criticisms are unacceptable. Efforts to retaliate against critics should lead oversight organizations to investigate possible misconduct by the retaliators (See the National Academies 2017 recommendations on fostering scientific integrity https://doi.org/10.17226/21896 accessed 1 June 2019.).

Second, consider the gold standard for strength of inference.

- Randomized, controlled experiment: Researchers should randomly select the subjects who receive treatment and those who receive no treatment (control conditions). Any departures from fully random selection should be documented and justified. If the treatment involves interventions that are presumed to have no effect in addition to the effective treatment, the control conditions should also include those interventions, e.g., placebo controls. Only the presumed effective component of the treatment should differ between treatment and control conditions, lest uncertainty about the effectiveness of treatment reduce the strength of inference.

- Lesser standards: We accept the rare need to study wild animals using the lower silver or bronze standards because some sociopolitical or biophysical settings preclude gold standard experiments. Such situational constraints should be rare. These lower standards lower confidence in the results by 50% or more. Silver standard or lower research is affected by uncontrolled factors that weaken inference. Many such factors can intrude. For example, the silver standard of before-and-after comparisons introduces the variable of time passing, because all subjects receive the treatment and its effect on subjects are followed over time. The bronze standard of correlational or observational study introduces many such potentially misleading factors because the researcher did not exert control over the intervention timing, magnitude, design, or the subjects receiving it.

- Higher standards: We defined the higher platinum standard that strengthens inference beyond the gold-standard of randomized, (placebo) controlled experiments without bias (see below for more on bias). The platinum standard includes both cross-over design and some level of blinding. Cross-over design requires the reversal of treatment and control within subjects. Because of randomization, some subjects will begin as placebo controls and others in treatment conditions, but all subjects will reverse to the other condition at approximately the same time midway through the experiment. A third reversal further strengthens inference about the effect of treatment. Blinding refers to concealing aspects of the experiment from different persons responsible for different portions of the research team or from reviewers, as we explain next.

- Single-, double-, triple-, or quadruple-blinding: Blinding is a design element intended to further reduce possible intentional or unintentional bias by researchers. The amount of blinding (single-, double-, triple-, or quadruple-) refers to how many steps in the experiment are concealed from researchers or reviewers. The steps that might be blinded include: (i) those intervening randomly should be unaware of subject histories and attributes and should not communicate which subjects received the control or treatment intervention to others in the research team (this depends on having used an undetectable intervention); (ii) those measuring the effects are unaware of which intervention the subject received (this too depends on having used an undetectable intervention); (iii) those interpreting results are unaware of which subjects received treatment or control; and (iv) those independently reviewing results are unaware of which subjects received treatment or control and unaware of the identity of the scientists who will or have conducted the research. Because science knows no authority, only evidence, blinding independent reviewers to conceal all unnecessary information might avoid several forms of bias (below). Note that blinding steps (ii) and (iii) might be feasibly done by the same set of people but the role in step (i) should be separate from all other roles to assure the success of blinding, and the role in step (iv) should be separate from all other roles to meet the criterion of independence.

Third, consider potential biases (intentionally or unintentionally slanting evidence to favor or disfavor one hypothesis or treatment) especially when it favors the scientist's preferred result.

- Selection or sampling bias and selecting a suitable sample of subjects: Any research on animals should consider the minimum number of subjects needed to detect an effect of intervention (treatment), while at the same time minimizing the infringement on the lives of those animals. Hence sample size is both a scientific and ethical decision that should be made transparently and subject to external review (see above). Once the appropriate number of subjects has been identified, selection of which subjects to investigate demands the utmost care to prevent self-selection bias and researcher bias, both of which might lead to treating subjects likely to show an effect of treatment. Self-selection and researcher bias are forms of selection bias that are very common and pernicious sources of unreliable findings. Random assignment is recommended to avoid the worst form of bias, which is systematic error in favor of a preferred results. When randomization is impossible, the next best procedure is blinding the selection and choosing subjects haphazardly without regard to their attributes or history and without regard to the potential effects of treatment or control.

- Treatment bias: This bias arises when placebo control or treatments are applied without regular, consistent intervention methods (e.g., haphazard doses of a medicine). The worst form of treatment bias is systematic for a favored result, when the timing, magnitude, or quality of the intervention is tailored to the history, attributes, or susceptibility of the subjects. Blinding (see above), standardized intervention protocols, and registered reports (see below) are reliable defenses against treatment bias.

- Measurement bias: This bias arises when measurement methods are inconsistent, imprecise, or inaccurate. The worst form of systematic bias arises when measurements are tailored to the history, attributes, or susceptibility of the subjects. Blinding (see above), standardized measurement protocols, and registered reports (see below) are reliable defenses against measurement bias.

- Reporting bias: This bias arises when analysis, of data, interpretation of results, or scientific communications misrepresent research methods or findings. The worst form arises when the reporting favors the scientists' preferred outcomes and naturally this is the most common form. Blinding (see above), standardized analysis protocols, and registered reports (see below) are reliable defenses against reporting bias.

- Independent review and publication bias: This bias arises when independent reviewers are favorably or unfavorably disposed toward the scientists, their results, or the nature of the scientific communications arising from the research. The worst form (and most common) arises when reviewers, editors, or publishers have an interest in findings or the power structures that might be affected by findings. A related form of independent review bias arises after scientific communications are made public, when critics try to silence or retaliate against the scientists who made those communications. Criticism should be welcomed but silencing or retaliating against scientists is unacceptable. The best defense against bias in independent reviews is the registered report and concealing the identity of authors from their peer reviewers. Registered reports are a new tool spreading in the scientific peer-reviewed journals. It adds an initial round of peer review of methods prior to data collection. If the first round of peer review accepts the methods, the journal commits to publish the findings regardless of the outcome, as long as no substantive changes in methods occurred after the first round of peer review. Registered reports guard against a publication bias that favors novel, striking results and disfavors confirmatory, replication efforts, while simultaneously guarding against reviewer bias that can favor or disfavor findings based on non-objective preferences of the reviewers.

I. Presuppositions

To set the stage for this document, we articulate our explicit pre-suppositions and starting points.

Science arises from a foundation of philosophy, ethics, values, and individual curiosities that originated long before the advent of animal research *per se*. Moreover, the modern practice of animal research is embedded in constitutions and the subordinate laws of many jurisdictions, which circumscribe how science is legitimately performed. Finally, within any particular tradition of science, there are domain-specific norms of scientific ethics, or how science is done to be accurate, precise, internally valid, predictive or generalizable, and persuasive to others. Given the latter foundations, this document should perforce limit itself to defined brackets of inquiry. Those brackets demarcate the scope of discussion without necessarily limiting the generalizability of our guidelines and recommendations beyond the brackets. We define those brackets next.

We presuppose that human decisions to act or not to act, and decisions how to act, should be based on (a) honest and thorough ethical deliberation [2], (b) on the legal protections afforded by constitutional provisions and the restrictions imposed by laws and regulations of the jurisdictions they intersect, and (c) on the best available information. We elaborate on the relationship between these three elements as interacting parts of a foundation for reliable evidence in Section II.

[2] (Lynn, 2018)

Regarding the ethics, we presuppose that ethics should not be separated from science. On the one hand, such separation risks immoral or unethical practices, and on the other, such separation might be impossible, because science is interlaced with values, ethics, and beliefs, some of which might be societal, and others might be personal. In our context, this interconnection of ethics and science plays out in the explicit integration of scientific integrity (the internal domain of ethics of scientific practice) into standards of evidence and into strong inference. We delimit our handling of animal ethics to questions of how animal ethics might affect scientific integrity and standards of evidence, not the many, varied and legitimate concerns about ethical treatment of animals that surround research but do not affect the standards of evidence so clearly. Even within our limited handling of animal ethics, several presuppositions should be identified immediately (Box 1).

Box 1. Some analogies between protections for human subjects and animal subjects

Past atrocities in research led to standards for the protection of human subjects in many countries. These vary in different countries but generally include moral principles of respect for persons, beneficence, and justice, which translate to specific practices such as informed consent and added protections for vulnerable subjects. A thorough summary is beyond our scope, but we broach the topic here to make two points. The first is that we presuppose a direct analogy to protection of animal subjects and to point out that some research, no matter how well designed, is ruled out on moral or ethical grounds. The same presupposition applies to animal research. We also presuppose that statements are not enough, processes should be accountable and proscriptions enforceable.

As of writing, there is broad disagreement whether protections for animal subjects should approximate those for humans or remain at their current, much lower levels of protection. We do not attempt to resolve this debate here, but instead we presuppose the philosophical approach of William Lynn [3], the ethical and regulatory approach of equitable consideration articulated by Santiago-Ávila [4] , and the practical and judicial approach of Just Preservation [5]. In brief, we presuppose that when research on animals is contemplated, scientists should engage in thorough, equitable, and explicit consideration of the interests of individual nonhumans as agents and subjects not automata and objects, whose interests are rarely if ever served by human intrusion. Although such research is likely to serve human interests primarily, we presuppose those human interests do not necessarily outweigh the interests of the individual nonhuman subjects. The articulation of this non-anthropocentric, pro-individual ethic, and a judicial

[3] (Lynn, 2015, 2018)
[4] (Santiago-Avila, Lynn, & Treves, 2018)
[5] (Treves, Santiago-Ávila, & Lynn, 2018)

process for deliberating on the use of wild nonhumans has been presented in full elsewhere [6].

Our presupposition that animals are not objects but subjects of research has been articulated well before, "People and many animals are 'agents' of their lives, responsible for some of their own behavior, respondent to the behavior of others, and not passive actors or mere subjects of ideologies, another agent's power, genetic code, or other causal forces." [7]. Other analogies also pertain. For example, vulnerable human populations – children, social minorities – are identified for special protection. We presuppose individual animals to be vulnerable subjects in research deserving of particular care and protection appropriate to their capabilities.

Further, we recommend that the process of consideration of nonhuman interests will include the latest evidence for that taxonomic group's capabilities for self-awareness, sentience, and socioemotional and cognitive abilities [8]. This is an example of where good scientific ethics (considering the existing literature) leads to better science and also promotes better ethics. Ethical conduct toward animal subjects cannot be limited to statements of concerns but should inform every step of the research process. Therefore, we alert the reader that our discussion of standards of evidence will point to ethical implications even if we lack the space to fully discuss these.

Regarding the law, we presuppose our audience is mainly subject to U.S. constitutional law and lower jurisdictional codes, although we note when possible if our presupposition can be broadened to other non-U.S. jurisdictions and whether certain codes or norms have been internationalized (Box 2).

Box 2: Does scientific integrity differ when researching free-ranging and captive animals?

We base the following arguments on U.S. public trust doctrine and Constitutional provisions, while noting that many other jurisdictions have differing provisions relating to animals, which might be stronger or weaker.

Since 1842 and explicitly in 1896, U.S. Supreme Courts declared wildlife to be part of the public trust [9]. This means that wild animals are not owned but rather

[6] (Treves, Santiago-Ávila, et al., 2018)
[7] (Lynn, 2018)
[8] Santiago-Avila, Lynn, & Treves, 2018)
[9] ("Geer v State of Connecticut," 1896; "Hughes v Oklahoma," 1979; "Martin v Waddell ", 1842): Hughes 1979 overturned Geer on the grounds of infringing the Commerce Clause but upheld its position on wildlife as a public trust responsibility. See (Blumm & Paulsen, 2013; Blumm et al., 2014; Blumm & Wood, 2017) for further details.

are part of a permanent trust for the sovereign U.S. public and future generations. Whether current and future generations are restricted to humans or include the futurity of all life on Earth remains an open legal question in our views [10]. Because wild animals are not private property, research on wildlife has a special character that seems legally different from research on captive animals.

Because captive animals may be owned as private property (even if wild-caught) and therefore in some ways protected from government intrusions by the U.S. Constitution's Property Clause, certain protections and privileges apply to the owners of captive, research animals. Without arguing whether these are ethically right or wrong, there are legitimate protections for owners of captive animals pertaining to privacy, trade, conveyance, management, etc. Such legal protections for private property owners imply some freedom from government taking and regulation – modulated by other Constitutional rights and assigned powers of the government. Therefore, scientists studying captive animals might not always be at liberty to be fully transparent (e.g., about animal procurement, handling, care, or disposal) about the methods used in captive research, or biased about the influence of captivity. For example, a scientist studying captive animals should even-handedly confront the many subtle limits on animals imposed by the conditions of captivity itself (e.g., lighting, position and actions of people, hardware), but the property aspect of such animals or constraints imposed by safety measures and captive conditions themselves might prevent the researcher from asking the fundamental paradigmatic question, 'is the effect of treatment caused by captivity or is it generalizable to other animals?'. To us, this seems to weaken the strength of inference when the effects of treatment include behavior or stress. Placebos should be designed with great care to avoid weakened inference, but some diminution of transparency might be unavoidable.

To be clear, we are not claiming that captive animals need less care. On the contrary, we believe they need more care because their freedoms and well-being are subject to additional liberties enjoyed by their owners, which rarely coincide with the interests of the animals. But the latter is an ethical question about animal care, while we are concerned with standards of evidence.

Nor do we suggest that research on captive animals implies less scientific integrity or produces research of lower-quality, but rather that we see a special character to captive animals as property that demands additional care when considering our guidelines for scientific integrity and strength of inference here. We leave it to subsequent legal experts and ethicists to determine if the

[10] (Treves, Santiago-Ávila, et al., 2018)

> standards of scientific integrity should differ for captive animal research and how.
>
> Finally, the U.S. is not alone in harboring constitutional and public trust provisions relating to people and animals [11], so we urge scientists of all nationalities adopt the strongest possible standards of evidence for animal research.

Finally, we presuppose that the best information to guide human action is that collected with the highest standards of scientific integrity. Individuals seeking information about nonhuman animals or about human interactions with nonhumans should be aware that their actions have ethical and legal implications. The evidence collected during their research, and how it was collected, will be scrutinized by many audiences having diverse views. Therefore, the quality of that information and the process by which it was collected and analyzed might influence the acceptance and the utility of the evidence in scholarship, practice, and policy.

II. Why standards are important to all research

The Brooks Institute for Animal Rights Law & Policy commissioned this document as a guide to all audiences evaluating animal research and as a means to articulate standards of evidence for those individuals and organizations conducting such research, and authorities and ethical committees responsible for issuing permits. This document is intended as a guide and therefore we must define our goals and terminology next.

Goals: In this document, we aim to describe guidelines and obligations of scientific integrity that support higher standards of evidence, particularly transparency, objectivity, reproducibility, and independent review, which we define in Section III. Second, we aim for standards of evidence that promote strong inference. We define strong inference and the research designs that produce it in Section IV. Within the scholarly and scientific fields that consider animal research, guidelines and standards that promote strong inference and scientific integrity should together speed progress towards new discovery of information that is reliable, accurate, precise, reproducible, generalizable, and persuasive to those seeking evidence. Therefore, as a third goal, we aim to identify steps in the recommended scientific process that might be changed in significant ways by through consideration of scientific integrity. The fourth and final goal is to provide a basis for systematic reviews of evidence that include past research, so the scientific community can proactively clean up the evidence base, consistent with the aims of reproducibility. Standards for cleaning up the scientific literature on animal research begin at Section V.

[11] (Blumm & Guthrie, 2012; Blumm & Wood, 2017; Treves, Artelle, et al., 2018)

Flexibility and rigidity in our guidelines and rules: Perforce, this document will evolve as needs and values change and discoveries are made about the scientific process [12]. Also, the constraints and obstacles are likely to change over time. The updates are presented at http://faculty.nelson.wisc.edu/treves/CCC.php. Therefore, we aspire to ideals (and methods) of transparency and strong inference, while acknowledging the difficulties imposed by current technologies and capabilities of scientists and the demands of scientific ethics (not animal ethics, a subject not treated here except superficially), all of which can sometimes erect seemingly insurmountable obstacles to strong inference. We acknowledge that well-intentioned scientists may have to make compromises in standards of evidence because they would adhere to the highest principles of scientific integrity. However, we are dubious about the converse (compromising scientific integrity or any ethical principles to achieve the highest standards of evidence), and therefore assert the principles of scientific integrity more forcefully while granting more flexibility in conforming to guidelines for standards of evidence.

Terminology, structure and organization of this document: We structure our discussion of standards of evidence around **scientific integrity** (the ethics of how one should collect and communicate **valid, reliable** and **generalizable** data) against a set of standards for **inference. Valid** means "well founded and fully applicable to the particular matter or circumstances; sound and to the point; against which no objection can fairly be brought" and **reliable** "able to be trusted; in which reliance or confidence may be placed; trustworthy, safe, sure" [13] refers to accurate measures or estimates (close to the true value) and precise estimates or measures (varying little from one to the next). **Generalizable** means using inductive logic from a specific case to a general set of cases. **Inference** is "the forming of a conclusion from data or premises, either by inductive or deductive methods; reasoning from something known or assumed to something else which follows from it" [14]. The latter general definition from the discipline of Logic appears free of criteria relating to the quality of inference. We define three types of inductive reasoning we use here and recommend methods for stronger inductive inference in Section IV, aimed at improving generalizability. Additional terms we use for process and subjects are **animal, data, scientist, research, inference, scientific communication, and independent review**.

- **Animal**: We focus on free-ranging, nonhuman vertebrates, whether descended from domesticated relatives or wild ones (hereafter wildlife for wild animals interchangeably). By free-ranging we refer to individual animals that are permanently or virtually always able to move where they choose in their ecosystem. We use captive to mean animals in captivity regardless of whether they or their recent ancestors were wild or domestic animals. We follow Hare in defining domestic(ated) to refer to animals whose ancestors were artificially selected by humans for tameness (lack of fear of humans or lack of aggression to humans), which often includes selective breeding [15]. Therefore, domestic animals might also be free-

[12] (Bernstein, 1991; Lynn, 2004)
[13] (OED, 2018)
[14] OED. (2018
[15] (Hare & Tomasello, 2005)

ranging if no human maintains control over that individual's reproduction or movements (sometimes called feral). In parallel, a wild animal might be captive. It does not become a domestic animal as a result of capture. A domestic animal on a farm that labels its activities "free-range" is still a captive animal unless it meets the definitions of free-ranging above. We focus on vertebrates because their independent agency is difficult to deny but note the possibility that many if not all of our guidelines might apply to other taxonomic groups, depending on the conclusions of research into agency self-awareness, sentience, capacity to suffer, socioemotional cognition, etc. Further specification of taxonomic boundaries is beyond our scope. Our focus on free-ranging animals reflects the primary author's focus, and in no way denies the importance of scientific integrity and standards of evidence for captive, domestic, or companion animal research. However, we acknowledge that captive animal research requires additional guidelines and rules that pertain to holding animals in captivity, which are not included in this document. Moreover, in the U.S. and perhaps other countries, the legal constraints on research differ for wildlife and other animals (Box 2).

- **Data, scientists, research and scientific communications**: **Data** are observations or measurements prior to analysis that pertain to animals. A **scientist** discovers new **data** or analyzes or interprets pre-existing **data**, whether that person is lay or formally trained as a scientist or other scholar. Activities of discovering, collecting, interpreting, or analyzing data constitute **research**. We limit our focus to research that might conceivably lead to intrusion or intervention on the lives of animals. Scientists, data, and research are broad and diffusely defined because we want to focus on the best science using standards of evidence, not on who is collecting data or what is being observed. **Scientific communications** relate to the scientists specifically conducting research who report in any form on their methods or the interpretations of their findings before or after research is conducted. Therefore, by our definitions, scientific communications are not by secondary sources repeating a scientist's interpretations.

- **Independent review** is a formal activity by one or more qualified third-parties with an arm's length relationship without conflict of interest relating to the scientists conducting the research or conflict of interest relating to their findings. Conflicts of interest relate mainly to financial or career advancement issues, not to differences of opinion. These might include anonymous peer reviewers, editors of journals, publishers, and other individuals convened to evaluate the methods before completion of research or the scientific communications following completion of research. Independent reviewers are put in a privileged position by being asked to recommend modifications or even recommend acceptance or rejection of proposed scientific communications. The privileged position of independent reviewers arises from being given access to some or all of the necessary information to evaluate any component of the research before it has been made public. That privilege comes with a responsibility to adhere to the demands of scientific integrity and to declare any possible conflicts of interest, because the outcomes contribute to the evidence base for the public, decision-makers, legal professionals, and other scientists. We devote Section III.d to independent review but also scatter guidelines for independent reviewers after every section of this document as **Guidelines for independent review**. Independent reviewers are also subject to the demands of scientific integrity because their privileged position implies a responsibility as gate-keepers for the scientific record. Therefore, independent reviewers have a duty of diligence and that duty extends to others after scientific communications are made public. We explore post-publication review and correction in Section V.

Emphasis and conceptual framework: Because our focus here is on standards of evidence, we circumscribe our handling of scientific integrity to the domain of scientific practice concerned with generating the best evidence, bracketing out the many ethical concerns about human subjects protection and research on animals. We emphasize quantitative, experimental research on animals. We do this for two reasons. First, interventions by humans into the lives and habitats of nonhumans have the potential for the greatest harm and therefore require more justification, part of which is the strength of inference one gains from such interventions. Second, we present reasoned arguments why straightforward observation or quantitative correlations of unmanipulated phenomena are shakier grounds for inference about cause-and-effect or unobserved mechanisms than are rigorous, unbiased experiments. Nevertheless, we acknowledge that the tighter the quantitative correlation between two measured phenomena, the closer scientists come to identifying causal mechanisms. It is beyond our scope here to summarize or settle the potentially intractable debate between subjectivist scholars and objectivists, or qualitative versus quantitative science. We adopt an interpretivist approach somewhere in the middle, rather than an axiomatic one at either extreme, namely that the social, personal, and historical context of research shape its final form, but there is a strong component of reality outside human perception, which is not imagined in the mind of the observer. Therefore, we accept that subjective, quantitative approaches can exercise their inferential powers closer to the phenomena they observe and thereby win validity in many cases, while more quantitative, systematic approaches encounter a trade-off by isolating phenomena for measurement they enhance reliability. We certainly conclude that quantitative science offers stronger inference about causal mechanisms, yet we warn that dismissing qualitative, subjective and observational approaches to scholarship would commit the error of discarding potentially valuable information without fair evaluation. Our view of the relationship between scientific integrity and strong inference is depicted schematically in Figure 1. Our view of the integration of information-gathering and ethical and legal reasoning is depicted schematically in Figure 2.

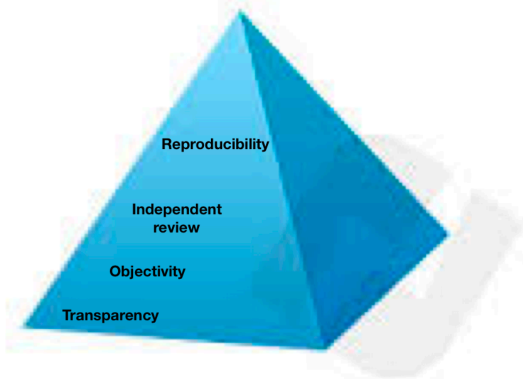

Figure 1. The pyramid of standards of evidence. One side is labeled with four principles of scientific integrity. All terms are defined in the text.

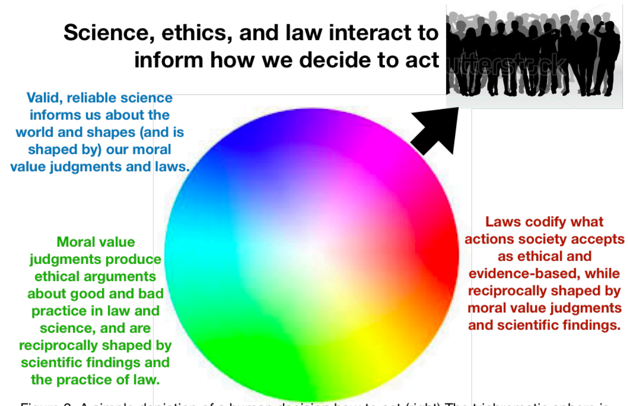

Figure 2. A simple depiction of a human decision how to act (right).The trichromatic sphere is intended to illustrate the interconnectedness of ethics, science, and law influencing each other (colors mixing). . This is one way of conceptualizing the relationship between ethics, law, and science

III. Scientific integrity

Because our focus here is on standards of evidence, we circumscribe our handling of scientific integrity to the domain of scientific practice concerned with generating the best evidence, bracketing out the many ethical concerns about human subjects protection and research on animals. Those and other ethical concerns are substantial and important, but beyond our scope.

Within the narrow domain of producing the best evidence, we define scientific integrity as research done with transparency, objectivity, reproducibility, and independent review. These principles are variously described and sometimes one or more are added or removed (see recent treatments in animal research [16]). We define these further below but first articulate the close relationship between them and the principles of strength of inference that follow. The strongest, possible inference might apply to evidence, yet it not prove actionable because of deficiencies in scientific integrity. For example, classified research facilities rarely subject their studies to independent review because of the top-secret clearances required. The evidence they produce might not pass muster as evidence for the purpose of legal arguments or public policy deliberations on the grounds of being irreproducible or lacking independent review. Conversely, the principles of scientific integrity might be followed to the utmost, yet a study might produce weak inference. For example, a hypothetical qualitative study of people's opinions of an animal species might generate weak inference, in the sense of yielding low confidence in its generalizability or predictive power (e.g., for attempts to predict how those respondents would behave towards that animal during a real encounter). Despite weak inference, the study might be perfectly transparent, objective, reproducible, and pass sterling independent review. Therefore, scientific integrity and strength of inference can be independent, yet they are related in many ways that we explain in Sections II–V and together combine to make what we believe are comprehensive standards of evidence.

[16] (Artelle et al., 2018; Carroll et al., 2017)

a. **Transparency**

Transparency is the continuous process and the habit of communicating clearly to make all information accessible to scrutiny. Transparency is usually essential to reproducibility, because independent investigators cannot reproduce research if its methods were not clear and complete. Also, transparency underlies independent review, because those scrutinizing methods, analysis, or interpretation can do a thorough job if no step is concealed. Transparency also supports objectivity ("the ability to consider or represent facts, information, etc., without being influenced by personal feelings or opinions; impartiality; detachment" [17]), because fair treatment of all observations and results of research demands that one expose potential conflicts of interest and biases at every step (Section III.a). We will also explain why transparency is essential behavior by independent reviewers (Section III.d). In sum, we treat transparency as the starting point for all science and perhaps the most important principle of scientific integrity, but not sufficient on its own.

Ut has long been known that methods should be disclosed transparently. But transparency about starting assumptions, such as value judgments, are seldom made explicit in our experience. Yet a scientist's value judgments about animals might increase or reduce bias (Bias is defined scientifically, in Section III.b on objectivity.). We address some of the more common unstated value judgments about animals that persist today.

We explore the vexing issue of value judgments that weaken inference or undermine scientific integrity by reference to financial conflicts of interest. Financial conflicts of interest, by definition, apply value to one outcome more than to others. It is widely accepted that a scientist who stands to gain personal or professionally from a certain research effort is vulnerable to unintentional or intentional bias [18]. The scientific community has long grappled with potential conflict of interest and has long required disclosures, so independent reviewers and audiences can weigh the scientific communication in that light of the scientist's potential to gain financially or by career advancement from particular findings. The same can apply to animals as instrumental goods for people or animals as property.

Most individuals value their own interests more than they value another individual's interests with notable exceptions of altruism, close kin, etc. Likewise, research subjects' interests usually differ from the scientists' interests in important ways. Animal subjects rarely benefit from research but institutions that conduct animal research for financial gain include wildlife agencies collecting permit fees for hunting, biomedical research centers patenting discoveries, and all individuals or organizations collecting overhead

---

[17] (OED, 2018)

[18] Pharmaceutical companies: doi.org/10.1136/bmj.326.7400.1167
Tobacco companies: https://read.dukeupress.edu/jhppl/article-pdf/21/3/515/358076/ddjhppl_21_3_515.pdf
Alcohol industry: https://doi.org/10.1136/bmj.326.7400.1167

from research grants. A scientist rarely benefits from questioning the claims of ownership, the conditions of captivity, the rights to hunt animals when the scientist depends on that organization for funding, access to data, access to animals, materials, staff, etc. Therefore, disclosure of property interests in animals could be as important as disclosure of traditional financial conflicts of interest.

Consider a hypothetical scientist who owns domestic cats that roam freely. Should that scientist disclose the property interest? The answer would be yes, for example, if they would benefit by discounting the potential, negative effects of free-roaming cats on birds. Alternately, the answer would be yes if their research addresses the costs and benefits of captivity, domestication, or free-ranging habits. Therefore, viewing animals as property, among other value judgments about animals, can affect objectivity and should be disclosed. Similarly, consider captive animals owned by a scientist or their organization. If an experiment were designed to evaluate if captivity affected tumor production, the owners of captive facilities would be considered to have a potential conflict of interest about the results, even if no explicit financial interest was at play. The inferences drawn from such an experiment would be weakened if the scientist(s) involved had a professional interest in the use of captive animals for tumor research. All of the above hypothetical scenarios expose how value judgments as fundamental as 'what I do to animals is OK' can affect objectivity and introduce bias that weakens inference. Therefore, we call for greater transparency than is traditional in animal research today.

Transparency by itself is unlikely to change bias. Yet, exposing a scientist's starting assumptions to view is an important first step for the benefit of independent review, all audiences for scientific communications, and in the best cases, for scientists themselves to become self-aware and take self-corrective steps.

Animal research is characterized by another pervasive value judgment that is usually implicit. That value judgment can be stated as 'humans take priority over nonhumans'.

Whether one agrees with anthropocentrism or not is not the point. Rather, we are making a claim about scientific integrity and inference. Value judgments are always implicit assumptions about relative value. Value judgments should be made transparent, when they affect our objectivity or the reproducibility of our science. Therefore, if anthropocentrism, non-anthropocentrism, and many other value judgments about animals, affect how one consciously or unconsciously assesses side-effects of research, how one chooses methods, and how one interprets results to favor our own interests over theirs. Value judgments about the interests of humans involved in research or likely to benefit from research *relative to the interests of the animal subjects* should be made explicit by scientists. But when and how?

In the examples that follow, we explore more specific cases of ownership or control of animals. Throughout, we focus on scientific integrity and strength of inference rather than the many important ethical issues that also arise.

Case study of anthropocentrism and non-anthropocentrism.

Value judgments about the priority of humans or animals and the importance of individuals over collectives such as populations or social groups can affect the inferences one draws from research. To illustrate the challenges, we summarize a case of three teams considering animal ethics in conservation [19].

The first team of authors advocated a soft anthropocentrism that granted priority for humans over nonhumans but advocated for more consideration of nonhuman interests in wildlife conservation decision-making. "While [our NA] principle is unequivocally non-anthropocentric, it prioritizes human well-being" p. 28, [20], but also "If a significant and genuine conservation interest calls for restricting a human interest, that restriction should occur except when doing so would result in injustice. When the restriction would be unjust every effort should be made by all involved parties to mitigate the restriction to the point of no longer being unjust." p. 30, [21]. By leaving open the definition of 'significant and genuine', whether social justice extends to nonhumans, and deploying the undefined term misanthropy elsewhere in their text, seemingly to anyone that disavows their soft anthropocentrism, the authors demonstrate anthropocentrism [22]. This in turn leads to a lack of objectivity about their recommendations for conservation ethics, which we describe after summarizing the second team's ethics.

The second team of ethicists and wildlife conservation scientists [23] "…promoted ecological justice distinct from both social justice and its "offshoot" (p. 369 [24]) of environmental justice, because these two schools of thought and practice express anthropocentric worldviews that … discount the interests of non-human nature." p. 138, [25]. They correctly exposed anthropocentrism masquerading as non-anthropocentrism which reveals how the work of the first team was less than fully transparent. Yet, when one reads the crux of their recommendations for reforming biodiversity conservation thought and practice, one finds a non-transparent value judgment about the priority of collectives over individual

[19] (Treves, Santiago-Ávila, et al., 2018)
[20] (J. A Vucetich et al., 2018)
[21] (J. A Vucetich et al., 2018)
[22] (Treves, Santiago-Ávila, et al., 2018)
[23] (Washington et al., 2018)
[24] (Washington et al., 2018)
[25] (Treves, Santiago-Ávila, et al., 2018)

animals [26]. For example, their final recommendation follows, "In particular, this assumes that populations, species and ecosystems have an interest in existing, persisting, maintaining, and re- generating their vital cycles, structures, functions and processes in evolution. It implies that conservation is no longer a process between people and about nature, but between nature and people, and justice has to be achieved between both." p. 372, [27].

This recommendation again "…seems to forget individual non-humans, which outweighs the scattered references to individuals throughout their text. We might view the above quotation as an oversight, but for its position in their final recommendations and too many signals throughout the paper suggesting the view collectives (populations, species, habitats, ecosystems) as more important than individuals..." p. 138, [28]. Although the second team of authors states a non-anthropocentric value judgment, they overlook the need to clearly articulate their preference for collectives over individual nonhumans. That lack of transparency compromises their objectivity in a similar fashion as the first team of authors as we explain next.

Any variety of ethical argument that does not recognize the interests of individual nonhumans in their own lives will fall short when individual human interests are pitted against individual nonhuman interests. Such happens "…often (all the time?) in conservation, because conflicts between individual humans and individual non-humans commonly face the rebuttal that 'the collective is not jeopardized by action x, so we can sacrifice the individual non-human [to gain] the benefits of action x'." p. 138, [29]. Consider a justice system in which an individual human can ask for the death of an individual nonhuman based on the justification that a non-lethal harm deserves a lethal response as long as only one individual nonhuman is killed. That would be unjust on its face, without knowing more about the harm.

Equitable treatment of individual humans and nonhumans might be unfamiliar or uncomfortable for some, but that itself should indicate the implicit action of value judgments. Such equitable considerations an objective starting point because it balances the preference of the actor (a human) against an opposing view that actor not getting its preference. That is why assumptions about anthropocentrism (weak or strong) and non-anthropocentrism must be more transparent in future animal research for the sake of scientific integrity.

[26] (Washington et al., 2018)
[27] (Washington et al., 2018)
[28] (Treves, Santiago-Ávila, et al., 2018)
[29] (Treves, Santiago-Ávila, et al., 2018)

Furthermore, such transparency can improve objectivity. The second team denied priority of humans but also implicitly prioritized ecological collectives (e.g., habitats or populations of nonhumans) over individual nonhumans. The third advocated non-anthropocentrism and consideration of individual interests in conservation decision-making [30]. Only the latter team exposed their own slightly different value judgments about our human responsibilities to individual nonhumans. That added bit of transparency about the authors’ own personal viewpoints might have improved their objectivity about an essential question in wildlife conservation: when is it appropriate for humans to kill a nonhuman?

The first set of authors’ soft anthropocentrism expressed concerns not to jeopardize the alliances and social relationships between conservationists and powerful interest groups. That concern is obviously irrelevant to animals subject to lethal interventions. Failure to acknowledge that preference for humans (and for current alliances of humans over future generations of people also) might make certain conservation interventions appear more effective than others at the outset. Such assumptions epitomize a loss of objectivity which can weaken inference about which conservation intervention would be objectively more effective. The second team eschewing such soft anthropocentrism instead prioritized collectives of nonhumans, which would at the outset make the putative conservation intervention of killing individual nonhumans appear more effective than a conservation intervention that protected individuals (e.g., non-lethal methods). Without an objective evaluation of two alternatives (kill individuals or don’t kill individuals) for a given problem, the scientists have diminished their own objectivity in preference for an unstated value judgment such as ‘we would protect the collective even if it means killing a nonhuman’. Pre-emptive dismissal of candidate actions or implicit preference for certain actions is by definition non-objective in scientific terms. The solution is not to dismiss one or the other, but to disclose the starting value judgment in both and let the inferences be drawn from there. Hopefully, the sunlight of transparency will resolve any issues affecting objectivity.

The author and the diverse peer reviewers who commented on this document differed on this question of how to disclose fundamental value judgments such as anthropocentrism. We do not try to resolve those differences here, but instead note them and recommend careful ethical and scientific examination. Below, we explore how to implement transparency in value judgments. We perceive a common bias towards human interests that place instrumental value on animals and judge the protection of animal subjects only through a lens of suffering or cruelty, not the broader recognition of animal interests as separate from human interests. Likewise, we perceive that the opposite bias is rarer but

[30] (Treves, Santiago-Ávila, et al., 2018)

also possible, wherein a scientist cares so deeply for animals that their interpretations of evidence are unduly influenced by compassion or empathy. Institutions or researchers who view animals only as property, as commodities, or amenities, etc. might lose objectivity just as much as animal researchers who view animals as sacred, in danger of extinction, or equal in capabilities to themselves. Therefore, transparency by research institutions and by the scientists themselves seems crucial to scientific integrity [31].

Certain types of animal research seem less likely to trigger potential conflicts of interest. Others are more likely, such as research leading to marketable products, research supporting the policies of powerful interest groups including government, and research that supports a certain political position. Yet, we do not have evidence to assert that concealing or disclosing a scientist's personal beliefs or actions toward animals (e.g., vegetarianism, hunting) would improve the strength of inference they or their independent reviewers might draw from research. Certainly, it is plausible that beliefs shape behavior, which in turn affect research as social psychologists have long demonstrated with correlational and experimental studies of attitude-behavior relationships [32], and more recently other social scientists have connected to social norms and peer group subcultures [33]. Yet, we lack a clear test for whether a scientist has insulated their personal value judgments from their interpretations of evidence. Those who claim such insulation is impossible have as much burden of proof as those who claim such insulation is routine, easy, or uncomplicated. We do not recommend an extreme position that scientists disclose every potential personal detail that might bias their interpretations of evidence. Rather we recommend the major well-understood biases that relate to personal profit professional advancement, be augmented by one or two disclosures of legal context and animal ethics, as we explain below.

Transparency about potentially biasing conflicts of interests and other value judgments are phrased below as questions that scientists undertaking animal research should answer explicitly for themselves initially and perhaps subsequently when making scientific communications that might be interpreted differently in light of the disclosures :

- Science is influenced by the laws under which it is conducted, and in turn, constitutions, laws, and regulations shed light on what is a potential conflict of interest. Different jurisdictions' constitutions articulate different protections for nature, and for their citizens' rights and governments' obligations to protect nature, sometimes including wild animals explicitly [34]. Therefore, one should ask, 'Do your national or local constitutions address animals or nature?' For one, the U.S. Constitution does not mention animals, but many world constitutions do [35]. The intent of this question is first

[31] (Wadman, 2018)
32 (Ajzen, 1991; Fishbein & Manfredo, 1992)
[33] (Kinzig et al., 2013)
[34] (Treves, Artelle, et al., 2018)
[35] (Treves, Artelle, et al., 2018)

to make scientists aware of animal law in their jurisdiction, and second, to reveal if certain scientific methods or outcomes might be more or less legitimate.

- Likewise, one might ask if a scientist feels ownership of animal subjects, which can apply even to a wildlife scientist: 'Does your animal research depend on animals that are private property or depend on skilled capture of animal subjects, such that you feel the subject animals have become your property?' The latter definition of captured wild animals as property is consistent with U.S. Supreme Court decisions dating back to 1896 or earlier [36]. Perceiving animal subjects as property might produce assumptions and privileges of property owners, compared to those perhaps felt by simple observers of animals. Whether wild or domestic animals are being studied, laws and constitutional provisions that relate to animals require disclosure. Similarly, the capture of animals and their subsequent release with equipment, such as transmitters, might impart a sense of ownership to the scientist or their allies; or does capture and instrumentation foster in some individuals a sense that scientists have liability and responsibility for the actions of those marked animals? As above, these questions might also have the salutary effect of encouraging scientists to understand the laws governing their activities. More importantly perhaps, this question gets at the perceptions of the scientist of animals as free-willed or inanimate property, a continuum of agency that might affect a scientist's objective choice of methods or interpretation of results.

- 'Do you or a collaborator in this research benefit from research on wild animals or from the findings?' For example, many wildlife scientists have experienced influences from hunting, trapping, hounding, angling, wildlife-viewing, agricultural interests, or protectionist groups [37]. For example, weakened inference have emerged from close ties to hunting interest, or alliances with domestic animal interests, e.g., academic peer-reviewed article titles such as "Why lions need to be hunted" [38], or in such articles, assertions such as, "The coexistence of a viable sheep and goat industry on American ranges and of large numbers of coyotes is difficult to visualize, and the sheep industry is dying for this reason" [39]. A scientist's formal affiliations to organizations that benefit from certain animal research findings – whether protective or exploitative – should be disclosed.

- 'Are you funded by or do you work for a government agency that promotes use of animals?' For example, many wildlife agencies espouse the North American model of wildlife management (NAM). The NAM was developed in the late 1970s and early 1980s as a mantra for wildlife professionals and then picked up by private and governmental

[36] P. 640, ("Geer v State of Connecticut," 1896)
[37] (Clark & Milloy, 2014)
[38] (Howard, 1988)
[39] (Shelton, 1973) —the empirical claim was later shown to be inaccurate (Berger, 2006) but we also present the quotation as evidence of a biased articulation of a research question.

associations of wildlife managers [40]'. The NAM has been scrutinized in great detail by scholars who found it introduced a bias towards hunting, angling, and trapping wild animals [41],. Furthermore, it lacks a solid legal foundation [42], and one analysis of hundreds of management plans for hunting terrestrial wildlife reveals a marked lack of the hallmarks of science despite its claims to be science-based [43]. Therefore, wildlife scientists should disclose if their professional advancements (grants, salary, or career opportunities) have benefited from a society or government agency that cites the NAM as a guiding principle. Other examples of organizations with implicit value judgments or explicit policies about animal research would require disclosure in the same vein. For example, a researcher with U.S. Department of Agriculture funding would have to acknowledge that government agency focuses on the instrumental value of animals primarily, and if the subdivision of that Department were Wildlife Services, the scientist should further disclose that their funding or advancement is associated with decades of practice and policies that prefer lethal management of wild animals almost to the exclusion of non-lethal methods [44].

- Data used by scientists but collected by employees of agencies that deal with human-wildlife conflicts might be biased when their measurements (e.g. determination of the culprit of a domestic animal loss) influence the decision about financial compensation paid to the private owner (e.g. they might tend to lay the blame on a protected wildlife species, which results in financial compensation, versus non-protected or domestic species, for which no compensation would be paid [45].

We are not attempting a futile effort to purge bias, only to expose it to the light (Box 3). We anticipate additional conflicts of interest will be detected and requests for their disclosure will ensue, but the guidelines above already go beyond what is commonly disclosed in scientific communications. Throughout this section on transparency, we emphasize the need for important disclosures. (The original author will follow the same rules below as an illustration of the approach after Box 3.) By demanding disclosure, we hope the guidelines will improve self-motivated efforts to neutralize bias, and most importantly, inform independent reviewers of possible influences on research design and the interpretation of findings.

We anticipate the rebuttal that institutions worldwide that conduct animal research have subjected scientists' proposals to independent review of protocols by boards well-versed in animal research. Setting aside debates about whether such checks on animal research are ethically or legally adequate [46], we are concerned here with scientific bias

[40] (Batcheller et al., 2010; Geist, Mahoney, & Organ, 2001; Prukop & Regan, 2005)
[41] (Clark & Milloy, 2014; J. A. Vucetich, J.T., Nelson, Peterson, & Bump, 2017)
[42] (Treves et al., 2017)
[43] (Artelle et al., 2018)
[44] (Robinson, 2005)
[45] (Plumer, Talvi, Männil, & Saarma, 2018)
[46] (Plous & Herzog, 2001)

resulting from non-transparency. Conflicts of interest have long been scrutinized by institutional review boards, because they might influence the harm to animals and the scientist’s evaluation of benefits-costs of research, Here we are more concerned with the weakening of scientific inference due to unstated value judgments.

Stating value judgments clearly, and transparently disclosing all such that relate to animals, should be celebrated and incentivized, not lead to jeopardy for scientists. We are aware of pitfalls of utmost transparency, in which the personal conduct of scientists towards animals is scrutinized and reprisals or sanctions follow. It would be unethical for independent reviewers or other audiences to discriminate against scientists for such transparency (Section V). The balance of transparency and legitimate privacy of scientists is not an easy balance and might need to be undertaken by institutions, editors, publishers, and employers charged with research integrity.

The need for utmost transparency might also be questioned ethically or legally, because of issues of personal privacy, but the scientist invoking such protections would bear the burden of proof. Voluntary transparency, e.g., via check-lists at the time of submission of scientific communications for publication, might not be adequate, as the following study revealed. For animals subject to intrusive research such as *in vivo* experiments, the road to transparent, thorough reporting and disclosure has been long and remains under debate. Since 2010, when the ARRIVE guidelines (Animal Research: Reporting of *In Vivo* Experiments) were first published, many journal editors and publishers modified instructions to authors with endorsement of the guidelines. In 2018, a multi-institution team reported on a randomized controlled trial that evaluated whether the scientific journal PLoS One accepted similar numbers of manuscripts that were “requested completion of an ARRIVE checklist” or followed “current standard practice” [47]. In a sample of 1689 manuscripts with 762 accepted, there was no significant difference in acceptance rate for the two groups of subject manuscripts. The authors concluded “The request to complete an ARRIVE checklist had no effect on full compliance with the ARRIVE guidelines. Details of animal husbandry (ARRIVE sub-item 9a) was the only item to show improved reporting, from 52.1% to 74.1%... These results suggest that other approaches are required to secure greater implementation of the ARRIVE guidelines.” abstract [48]. We interpret these results as a recommendation for journals to enforce, not only endorse, completion of the checklists. We are encouraged to note that PLoS One took a similar conclusion to heart in a 2018 blog post suggesting reproducibility of research would be improved by their intervention to foster transparency [49]. Checklists might be inadequate for encouraging transparent reporting on compliance with recommendations for animal research. That emphasizes that one goal of this document (articulating recommendations) is only a first step.

[47] (Hair, MacLeod, & Sena, 2018)
[48] (Hair et al., 2018)
[49] (Alvino & PLOS ONE Editors, 2018)

Box 3: Guidelines for transparency in animal research

All scientists conducting animal research should begin with a clear statement of whether they espouse a value judgment that humans take priority over nonhumans, or another speciesist worldview (speciesist indicates that individuals are valued solely by their species or other subgroup membership rather than equitably considered [50]).

Furthermore, scientists should disclose any potential conflicts of interest relating to personal profit, professional advancement, or other rewards for particular findings in their animal research. These conflicts of interest arise when a scientist, group, or government agency views animals as property, disposable, inferior, or only instrumental to humans. Because animal research can generate profit or political influence, scientists should disclose more rather than less, given that self-dealing or self-interest is antithetical to objectivity (see Section III.b).

Disclosure explicitly implies scrutiny. That scrutiny begins with self-awareness to set aside and expose personal value judgments. But disclosure is equally or more important to thorough independent review.

Guidelines for independent review: Independent reviewers should be the short-term judges of the quality of reasoned evidence, but they might not be able or qualified to pierce non-transparency. After all, scientific communications present only what the scientist wishes to share about research. Therefore, we call for more collaboration between overseers of scientific integrity and overseers of strength of inference.

Ultimately, the broad scientific community, current and future generations of the public, legal experts, governments, philosophers, and ethicists will render final judgments on whether bias interfered with evidence. Symmetrically, independent reviewers should also be transparent about their biases, which might precede a particular piece of research they are reviewing or might be triggered by the style or the transparent disclosures we recommend above of a particular scientific communication once placed in front of them.

Scientific journals and editors should require independent reviewers to answer the same questions as scientists did, although we acknowledge those disclosures might remain confidential in the hands of the editors [51]. Because anonymity of

[50] (Santiago-Avila, Lynn, et al., 2018)
[51] See for example such a growing database of financial disclosures for medical scientists at http://www.convey.org/

independent reviewers can be important to their objectivity (Section III.d), we recommend a limited transparency for anonymous reviewers. Namely, independent reviewers should respond explicitly to the same questions we recommend for scientists, but those responses might be stored confidentially with the editor or donor or published in aggregate to protect anonymity. We acknowledge that anonymity seems to fly in the face of transparency and we acknowledge a current lively debate on whether peer review should be anonymous in scientific journals. However, without evidence on the consequences of anonymity or identifying peer reviewers, we choose to present a guideline for the status quo, until a change is required by evidence. Certainly, more editorial oversight of the biases among peer reviewers seems important.

The scientists submitting animal research proposals and scientific communications should be obligated to answer the above questions in detail (not just checking boxes) and we recommend editorial or funder oversight of such statements. Furthermore, editorial decisions to reject, modify, or omit portions of scientific communications that relate to animal research seem to deserve transparent explanations also, lest editorial value judgments be imposed upon those who submit scientific communications for independent review.

Example of disclosing value judgments about animals: To illustrate the use of the transparency guidelines we have discussed above; the original author of this document will follow them and disclose the potential biases that affect this work. He held an implicit, soft anthropocentric worldview in animal research from 1991–2017, then a non-anthropocentric worldview 2017–present. He works for a large, state land-grant university that profits from animal research. His own research is on wildlife and his professional advancement has depended on that research including scientific communications to all interested audiences haphazardly before 2011 and following his own sense of duty as a public trustee since 2011. He currently has neither financial or other economic interests in harming animals nor in protecting animals. The commission of the Brooks Institute might be seen as a potential conflict of interest except that contract is transparent about the absence of influence over the content of this document. He serves as an unpaid science advisor to several non-profit groups that work to preserve nature, engage in moral or political discourse on animals, or produce scientific reports on wildlife. He declares that he does not espouse constitutional rights for animals but does espouse a view that nonhumans deserve equitable representation in human policy or legal proceedings when the interests of those nonhumans would be infringed by human decisions to preserve or use nature.

We recommend that specific institutions in a disclosure like the one above NOT be named unless they have a direct financial interesting he outcome and the scientists have positive or negative relationships to those organizations.

Transparency is not enough. Because some biases are so powerful and pervasive, transparency alone is not sufficient to nullify bias and produce strong inference. Strong inference also requires objectivity, reproducibility, and independent review to nullify a paradigmatic bias. Therefore, we turn to objectivity next.

b. Objectivity

We follow the Oxford English Dictionary's definition of objectivity from philosophy as, "the ability to consider or represent facts, information, etc., without being influenced by personal feelings or opinions; impartiality; detachment" [52]. In the previous section on transparency, we explained how individual value judgments and paradigmatic worldviews of animals, and professional or organizational conflicts of interest might weaken inference and should therefore be disclosed. We define bias in preparation for our argument that transparently objective scientists can reduce the bias imposed by paradigms or assumptions (Box 4).

Box 4: Bias defined scientifically

We define bias in its scientific sense to mean systematic error as opposed to random error or as in statistics, "A systematic distortion of an expected statistical result due to a factor not allowed for in its derivation; also, a tendency to produce such distortion"[53]. All measurement involves error due to the imperfections of instruments and human beings handling those instruments.

Much of measurement error is random, because it is equally likely to under- or over-estimate the parameter we are measuring. Random error or error that is haphazard with regard to the hypotheses we are testing or the research questions we are answering tends to be relatively harmless by reducing confidence in the answer not generating an invalid answer, as long as the magnitude of that error is not great relative to the precision of the answer one seeks.

However, bias is systematic error – a slant or a consistent error – in favor of one direction, one outcome, or one result over others. That understanding of bias in science also fits a classic definition of bias that we believe can apply to qualitative science, "A swaying influence, impulse, or weight; 'any thing which turns a man [sic] to a particular course or gives the direction to his measures.'" [54]. Again, here we aim for a middle ground between assertions that objective science is value-free on one extreme, or assertions that all science is equally biased by personal perspective on the other. Because all research arises from a

[52] (OED, 2018)
[53] (OED, 2018)
[54] (OED, 2018)

human perspective (in the sense that all measurement or observation is done by humans or human tools), the scientist's perspective is always present. Yet, bias as defined above can be partially overcome by the careful application of scientific integrity and standards of evidence while insulating the measurement or observation from insidious, systematic error.

Systematic error prevents us from making accurate estimates (close to the true value) – even though they might be very precise (varying little one from the next estimate). Bias can arise from defective instrumentation or from human behavior. When bias emerges prior to measurement (sampling bias or treatment bias, Sections IV.a and b), or after measurement during analysis and reporting (measurement bias, reporting bias, or publication bias, Sections IV.d and e, V), then the accuracy and precision of measurement is likely to be irrelevant because inference is weakened by biased design, analysis, or interpretation.

The tricky aspect of bias is that every human being has it. Bias emerges in the questions a scientist asks and how they go about answering those. Objective scientists learn how to set most of those preferences aside as much as they are aware of them and making their biases as transparent as possible, when communicating about the research.

Two guidelines for preserving objectivity and diminishing bias come from philosophers of science. Dr. A. Gawande wrote elegantly about skepticism, citing Dr. Edwin Hubble among others [55], when he explained how scientists think. They are chronically skeptical of ideas, even their own, and they know knowledge is always vulnerable to new evidence because our understanding of reality is just a process of progressive approximation. Also, long ago, Dr. T.C. Chamberlin recommended the procedure of multiple working hypotheses for avoiding an undeserved preference for one hypothesis over another [56]. Failure to be aware of one's own preferences for certain questions or certain answers in research can introduce bias into one's interpretations and scientific communications.

Cooperative research, in which collaborators are able to question each other, scrutinize all aspects of the research, and revise before reporting, might improve transparency about the biases that remain and attain greater objectivity than a single scientist can attain. Collaboration, like independent review, seems to reduce bias because virtually no two persons share identical biases at every step in research and scientific communication. This might help to explain why natural

[55] (Gawande, 2016)
[56] (Chamberlin 1890)

> sciences have tended towards larger teams of collaborators over time, as awareness of unintentional insinuation of individual bias has grown.

To increase objectivity in animal research, we recommend all three of the above remedies for bias: skepticism, the method of multiple working hypotheses, and collaborative research.

Paradigms also produce bias that is more pervasive and difficult to diminish. Paradigms are worldviews and a scientific paradigm is, "A conceptual or methodological model underlying the theories and practices of a science or discipline at a particular time; (hence) a generally accepted world view." [57] Paradigms shape how one views the universe, the types of questions one asks, and the interpretation of the answers one gets. These are value judgments that should be disclosed transparently as in the previous section. However, we suspect paradigms will rarely be disclosed properly because their holders are often unaware of them, and reviewers might share the same beliefs. Therefore, in paradigms we have a persistent, systematic bias that might affect every step of the scientific process. We emphasize the word 'might' here because outstandingly objective scientists can surmount the prevailing paradigms of their time.

Our duty here is to discuss how one might constructively address paradigms in animal research. We approach this question and more mundane gaps in objectivity below and in the following section on reproducibility. Yet, we acknowledge that some beliefs or worldviews about animals and about animal research are difficult to set aside or even to falsify in one or even many lifetimes of scientists. For example, the various beliefs in a 'balance of nature' including natural theology, Aristotelian teleology, and creationism have been deeply influential in slowing progress in ecology, among other fields. For example, the various paradigms of teleology – that nature is essentially determined, not forever changing through an unplanned process of adaptation to local environments combined with non-equilibria continuous perturbation [58] – held sway for >2000 years because it fit individual scientists' beliefs about god(s) creating the universe. Some paradigms are long-lasting yet falsifiable by careful observation and experimentation (e.g., the gene as the sole unit of heredity [59]), whereas others such as a belief in the balance in nature [60] took thousands of years to falsify and still resist efforts to discard it.

In animal research, we see two widespread paradigms about whether humans can grasp and adequately represent a nonhuman's experience or not. These paradigmatic worldviews might be particularly fraught in studies of cognition, intention, and cognitive ethology [61]. Disclosing these assumptions transparently is a good start, but if the reviewers and most readers share them, we need another step to achieve the strongest inference. We need an objective statement of the assumption and its alternative including the consequences if the alternative is

[57] OED, 2018)
[58] (Kricher, 2009)
[59] (Durham, 1991; Mukherjee, 2016)
[60] (Kricher, 2009)
[61] (Bekoff, 1996; Treves & Pizzagalli, 2002)

found to be a better representation of reality. Then we need a reproducible experiment to distinguish the two opposed assumptions, which calls for conceptual reproducibility (Section III.c).

To understand better how a paradigm can be over-turned even while a scientist is embracing it, we should consider how paradigms have been over-turned in the past. One view holds that paradigms are deep-seated beliefs held by individual scientists and shared by scientific communities [62], which are almost never overcome in one lifetime, except by singular scientific geniuses (e.g., Einstein, Darwin). This hypothesis predicts that paradigms change when new generations of scientists with different belief systems begin to observe nature. The second view of paradigms is that they are identical to hypotheses but simply harder to falsify. This view predicts that paradigms persist until exposed to strong inference. An essential starting point for strong inference (Section IV) is opposed hypotheses. As far back as 1890, Chamberlin [63] articulated a method for stimulating objectivity and protecting it throughout the scientific process. His method of multiple working hypotheses was visionary, "With this method the dangers of parental affection for a favorite theory can be circumvented" [64], because it encouraged scientists to recognize when they became fond of one hypothesis with its attendant loss of objectivity. We add that a priori identification of all hypotheses is a more objective approach than a posteriori articulation.

Indeed, Platt[65] took Chamberlain's ideas further to identify pitfalls in slow-moving fields of science. He blamed slow progress on an inability to oppose hypotheses authentically (i.e., straw men hypotheses instead of reasonable alternatives) or an inability to design experiments that would test between the opposed reasonable hypotheses. He blamed excessively complex models including too many variables and excuses about the infeasibility of incisive experiments. Following these traditions, recent work has also called out unfalsifiable claims that objective experiments cannot be conducted on wild animals [66]. Platt [67] hypothesized that scientific disciplines advance quickly when they elegantly oppose contrasting hypotheses and design incisive experiments to distinguish the opposed hypotheses. In Platt's view, a paradigm can be falsified if it is honestly opposed to an alternative and an incisive experiment is designed and run successfully. Chamberlin's and Platt's processes require objectivity and conceptual reproducibility, not a genius.

Yet, paradigms persist, probably because of ethical and legal constraints and presuppositions that reflect unshakable worldviews of the universe that reinforce allied paradigms of science. Nevertheless, we believe that when a scientist can objectively evaluate evidence and

[62] (Kuhn, 1962)
[63] (Chamberlin 1890)
[64] P. 754, (Chamberlin 1890)
[65] (Platt, 1964)
[66] (Ohrens, Bonacic, & Treves, 2019; Treves, Krofel, & McManus, 2016)
[67] (Platt, 1964)

objectively design research, the probability of strong inference rises because the data will tell the story rather than the story being predetermined by the biases and paradigms at hand.

Box 5: Rules for objectivity in animal research

The principle of objectivity demands that scientists nourish skepticism about their own results, support reasoned, alternative hypotheses throughout the research process, counter with reasonable, alternative hypotheses before drawing inferences, and embed their scientific communications with cautions about the limits to their inferences.

Animal research rarely addresses broad, paradigmatic questions, but when it does, the utmost transparency and objectivity is required to oppose alternative paradigms. In such cases, conceptual reproducibility of findings using an alternative paradigm appears essential (Section III.c).

Most animal research addresses alternative hypotheses within the same paradigm. Objectivity demands a sober look at how to evaluate the alternatives fairly. Furthermore, if alternative methods to reach the same outcome are available, an objective scientist will choose the more transparent and reproducible method. Also, when scientists make a judgment to prefer certain methods, emphasize certain results, or interpret analyses in a particular way, it is better to make such judgments on reproducible and objective criteria not on experience, authority, or professional position (criteria that are generally not objective or reproducible).

When alternative hypotheses are being opposed, the objective scientist will stack the odds against their preferred hypothesis. The attributes of objective scientists include humility to acknowledge that the alternative might be equally reasonable, a willingness to invite and ponder criticism and correction, and a preference for debate and replication rather than defensiveness against critique or independent replication efforts (Section V).

Guidelines for independent review: The same rules for objectivity of scientists apply to objectivity of independent reviewers. It is not unusual – and strikes us as a good practice – for reviewers to cite evidence – whether contrary or corroborating – when evaluating new evidence before them. This practice is in opposition to simply asserting that the new evidence is contradicted or corroborated by prior information.

If the process for presentation of contrary or corroborating evidence was itself not consistent with scientific integrity, then the corroboration or contradiction should not be considered definitive. In short, independent reviewers are dutybound to sift and winnow evidence objectively, just as scientists are obligated to do so. Finally, we recommend reviewers identify their paradigms as clearly as they can to themselves at very least, and possibly to the editors or overseers, if review is anonymous. Editors should also be governed by rules of transparency and objectivity about their decisions to assign independent reviewers' or make decisions on scientific communications in front of them (also see Section V on when to correct or retract published scientific communications).

c. **Reproducibility**

We define reproducibility as the quality of a scientific finding that can be replicated by other scientists under the same conditions as the original. Reproducibility is fundamental to science, because falsifying inadequate hypotheses is essential to closer and closer approximations of reality [68]. If an explanation cannot be disproven (falsified), it is not scientific. Therefore, falsifiability demands reproducibility of inferences and research findings (often referred to as replicability). This hallmark of science is as true for large-scale theories of how nature works, as for small-scale explanations of results in animal research.

Recent studies indicate that the current scientific record suffers from a marked lack of reproducibility, because over 70% of efforts at replication failed to replicate the original findings [69]. This grim finding has been tempered somewhat in 2018 by a replication effort aimed at testing the hypothesis that studies in psychology fail to replicate because different populations of humans were being tested during replications. The new study concluded that 14 of 28 classic and contemporary psychology studies were replicated at 60 different labs in 36 countries [70]. Because behavioral psychology has been the most forthright and exposed to the reproducibility crisis in popular media reports, the latter study corroborates those who said the so-called crisis of reproducibility in science has been exaggerated, while at the same time, offering fresh empirical evidence to address the remaining concerns about the high proportion of published studies that cannot be replicated [71].

We assume that if an independent, qualified scientist working in good faith cannot replicate a finding, there is a problem with the original. If a hypothesis is falsified and independent scientists do so reproducibly, another hypothesis is needed. If the hypothesis that is falsified in the latter manner is fundamental to a particular scientific discipline, a new paradigm might be needed.

Here we define three categories of reproducibility. The first is exact reproducibility or the replication of results when every step, condition, and material can be replicated. Exact

[68] (Popper, 1959)
[69] (Baker & Brandon, 2016)
[70] (Klein, preprint 2018)
[71] (Alvino & PLOS ONE Editors, 2018; Baker & Brandon, 2016; Goodman, Fanelli, & Ioannidis, 2016; Ioannidis, 2005; Kretser et al., 2019; Open Science Collaboration, 2015)

reproducibility is almost never achievable in animal research because time has passed, or conditions vary by location or individual animals. We include it here because some animal research is conducted within artificial systems (e.g., computer models) that should in principle be exactly reproducible. Indeed, exact reproducibility depends on a trade-off with verisimilitude. However, given the rarity of exact reproducibility in animal research, we focus more attention on technical and conceptual reproducibility below.

Technical reproducibility is defined as following the same methods to replicate a result, without requiring identical times, locations, materials, investigators, animal subjects, etc. [72] This differs from conceptual reproducibility that attempts to reach the same result by a different pathway (Figure 3). Conceptual reproducibility is essential to test opposing paradigms in science, wherein the alternative pathway is the alternative paradigm (Section III.a). Because exact and technical reproducibility should adopt the starting paradigm, but conceptual reproducibility need not do so, the two sets of approaches are complementary.

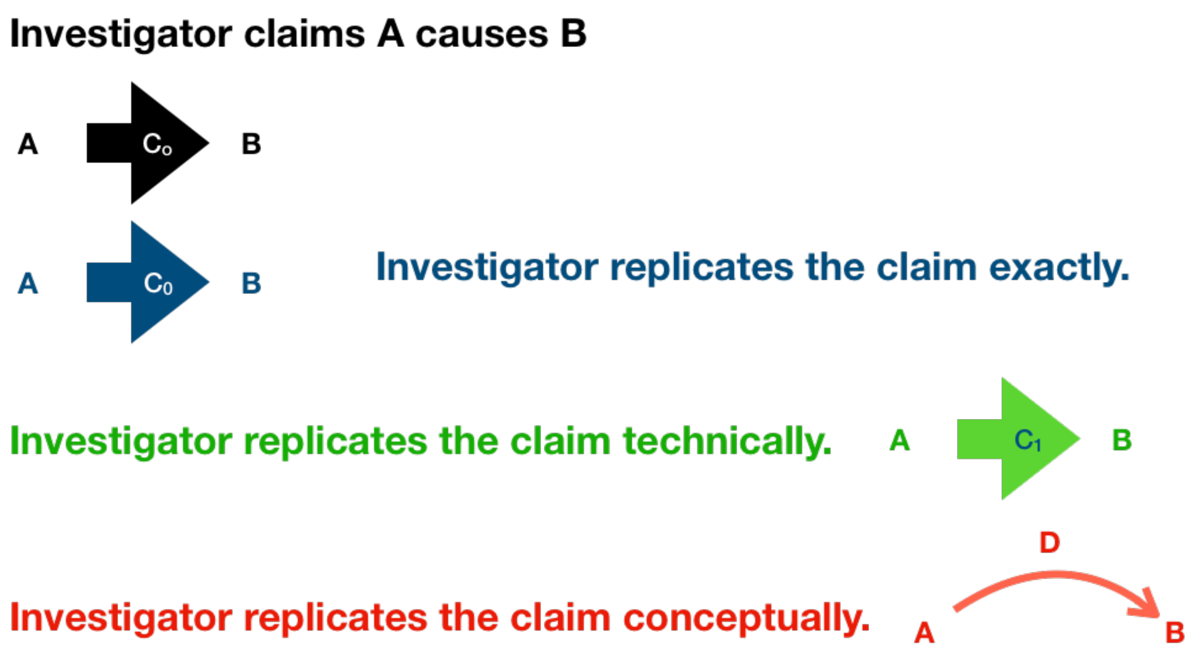

Figure 3. Reproducibility of three types to produce effect B. Exact reproducibility (blue) arises when an investigator uses identical methods (C_0) including the same times, locations, materials, subject animals, etc. to find B. Technical reproducibility (green) is less stringent because it allows independent researchers to use the same methods (C_1) without demanding identical conditions. Conceptual reproducibility (red) uses a different pathway (D) to generate the same conclusion that A causes B.

[72] (Goodman et al., 2016)

All forms of reproducibility are important because one can be fooled by technical reproducibility into repeating an error, whereas conceptual reproducibility, although rarer, tests the strength of inference. An example from cancer research illustrated how two entangled biochemical pathways had to be disentangled through an exercise of conceptual reproducibility because repeated efforts at technical reproducibility had simply confirmed the wrong pathway yet failed to explain new evidence about the etiology of the cancer [73]. Likewise, the 20th century history of cell cultures contaminated by Henrietta Lacks' preserved cancer cell line reveals the pernicious effect of a persistent and erroneous assumption repeated time and time again in exercises of technical reproducibility [74]. In those cases, Ms. Lacks' cancer cells had contaminated other cell lines that were assumed to be pure, and countless captive animals were wastefully used in flawed experiments and human patients were misled, because decades passed before anyone undertook the exercise of conceptual reproducibility to test if other cell lines had been contaminated by Ms. Lacks' immortal cell line. Returning to the balance of nature, Kricher [75] describes how paradigms of plant species assemblages were subject to falsifiable tests of both Clements' (1874-1945) older ideas of orderly succession and Gleason's (1882–1975) newer ideas of stochastic processes that varied from site to site. The rejection of the former paradigm forever, reflects the superior inferences about how any plant assemblage changes over time based on testable predictions about migration, competition, and adaptation to local conditions rather than on unfalsifiable claims about essential attributes of particular species and orderly change from one predetermined state to another. Likewise, Darwin's paradigm of descent with modification of animals and other organisms offered the new mechanism of natural selection [76], which replaced various, earlier paradigms of essentialism or acquired characteristics [77]. Meanwhile Darwin's paradigm of heredity was wrong [78] and it was replaced later by the Mendelian paradigm in an often-told story about peas [79]. That paradigm too has made room for a newer paradigm with the discovery of epigenetics [80]. In sum, the strongest scientific inferences arise when one transparently exposes one's biases, paradigms, and assumptions to objective scrutiny and to falsification through conceptual reproducibility. In Box 6, we offer an example of conceptual reproducibility in animal research and how this principle generally recommends a non-lethal approach to animal research.

Box 6. Why scientific integrity favors non-lethal animal research

Here we argue that the principles of objectivity and reproducibility both favor non-lethal methods of intervention over lethal methods in animal research.

[73] (Pairs, 2018)
[74] (Skloot, 2010)
[75] (Kricher, 2009)
[76] (Darwin, 1859)
[77] (Kricher, 2009)
[78] (Mukherjee, 2016)
[79] (Mukherjee, 2016)
[80] (Mukherjee, 2016)

Less harmful, indeed non-lethal methods, would generally be preferable on scientific grounds. For example, the traditional and most common method today for preventing wildlife damage to property is to kill the suspected culprit animals. This approach actually seeks to protect property (the end goal), rather than seeking to kill an animal (the ostensible means to that end). Therefore, any intervention that makes it impossible or nearly impossible for the suspected culprits to damage property would achieve the desired goal. If such an alternative to killing would do less harm, the alternative non-lethal intervention should be preferred ethically and legally. It turns out, the less harmful and preferably non-lethal method is also preferable scientifically as we explain next.

The reason that non-lethal methods of animal research should generally be preferred scientifically is because of several principles of scientific integrity. First, in the specific case of predation on domestic animals, the peer-reviewed, published, scientific literature on effectiveness of non-lethal methods has yielded stronger inference than the evidence for lethal methods [81]. Second, we expect thorough transparency is better served by a living animal that can be examined and its life functions measured than by a dead animal. Third, by the principle of objectivity, one should consider the most precise statement of cause and effect over less precise ones. In this case, effect y might be reached by preventing a behavior, rather than preventing all behaviors. Fourth, and most importantly, one should prefer the more reproducible of two approaches. If a scientist kills an animal to test if x causes y, then exact reproducibility will never be attained, and technical reproducibility will demand yet more killing. A pathway that allows exact and conceptual reproducibility provides stronger inference, i.e., do not destroy your materials.

Scientists who embrace the idea of conceptual reproducibility and the need for less-harmful, non-lethal approaches to animal research should beware of the following false choice. Just because non-lethal methods provide conceptual reproducibility to strengthen inferences made from prior lethal research does not logically require that non-lethal methods be conceptually reproduced with lethal methods. That would be a false choice because the inference is about the effect of an intervention, not its mechanism. Therefore, there are usually several non-lethal methods to achieve the same outcome. Each link in the hypothesis of cause-and-effect deserves a test and replication, but none require killing animals. The only scientific reason to kill a research animal is to measure something that can only be measured once it is dead.

[81] (Eklund, López-Bao, Tourani, Chapron, & Frank, 2017; Lennox, Gallagher, Ritchie, & Cooke, 2018; Miller et al., 2016; Moreira-Arce, Ugarte, Zorondo-Rodríguez, & Simonetti, 2018; Treves et al., 2016; Lily M. van Eeden et al., 2018; Lily M. van Eeden et al., 2018)

Traditionally, wildlife managers assume that problem animals exist, once they start damaging property they ostensibly would not stop, and problem animals would transfer such behavior to associates or offspring. If one proposes a non-lethal method to stop an animal from damaging property and that method fails, some traditionalists would argue that animal should be killed. That is a false choice. Another non-lethal method can be tried, as above for reproducibility. Furthermore, the hypothesis that the problem animal will repeat the behavior has not been shown experimentally. Nor has the hypothesis of horizontal or vertical transfer to associates been demonstrated experimentally. In turn, if associates adopt the behavior, one can test the non-lethal deterrent on them as well, i.e., teaching aversion just as damage to property was putatively taught. Furthermore, if non-lethal behavioral modification (e.g., aversive conditioning) is successful, this can be transferred to other individuals. On the other hand, conflict-prone locations can repeatedly promote conflict behavior in new animals that replace those that were killed or removed, thus perpetuating the conflict despite intervention.

In sum, principles of scientific integrity prioritize less-harmful and non-lethal methods of animal research. However, our discussion above should not be construed as a rationale not to evaluate lethal methods experimentally. If an actor is using lethal methods, as many governments and private individuals do today, then scientific evidence is needed to evaluate if such lethal methods are indeed functionally effective in protecting human interests. We have articulated this need repeatedly on scientific, legal, and ethical grounds [82]. Our larger point is that if the goal is property protection by the most effective method, then non-lethal methods should be evaluated first and exhaustively first.

Conceptual reproducibility strengthens inference more than technical reproducibility as the preceding examples and Box 6 elaborates. In brief, when a cause-and-effect relationship is confirmed by two independent pathways, the probability that scientists have been misled by confounding variables, biases, or logical fallacies, diminishes (Box 7).

Box 7: Guidelines for reproducibility in animal research

Transparent and objective methods are the essential starting points for reproducibility. Collegial and constructive collaboration with independent scientists wishing to replicate a finding should be undertaken (Section V). Furthermore, until a piece of evidence has been reproduced conceptually, at

[82] (Santiago-Avila, Cornman, & Treves, 2018; Santiago-Avila, Lynn, et al., 2018; Treves, 2009; Treves & Bruskotter, 2014; Treves et al., 2017; Treves et al., 2016; Treves & Naughton-Treves, 2005; Treves, Santiago-Ávila, et al., 2018; Lily M. van Eeden et al., 2018)

least once independently, it should be considered tentative, no matter how strong the inference appears to be in the original result.

Scientists planning animal research should consider carefully the issue of future reproducibility, because some animal research entails death or irreparable harm to animals. Exact reproducibility will be impossible in such cases and even technical reproducibility might be compromised. Animal research that involves serious harm to animals demands that one first use methods that are least harmful. If harmful methods were used in the past, less-harmful methods would provide conceptual reproducibility which is an important asset in strengthening inference. If non-lethal methods came first, it is not necessary to use lethal methods for conceptual reproducibility.

Guidelines for independent review: The principle of reproducibility demands that independent reviewers ask the authors of scientific communications for meticulous detail in methods. Omissions of information are both violations of transparency and of reproducibility. Basically, an independent reviewer should conclude they might replicate the study if sufficient time and resources were available (see various practical examples of how replication efforts are coordinated [83]).

Independent reviewers of proposals to harm animals should scrutinize the justifications for why a less-harmful method will not be attempted. Justifications based on feasibility are not generally valid, while justifications based on impossibility should also be scrutinized closely in case the convenience of the scientists conducting the research is the argument for impossibility (Box 10).

The principle of reproducibility puts a burden on independent reviewers also. Independent reviewers should not cite evidence that is itself irreproducible when evaluating new evidence before them. Perhaps more common is to raise claims that are unfalsifiable in support or opposition to new evidence. For example, the claim that randomized experiments cannot be run in systems of predators and domestic animals and that domestic animal owners will not accept the placebo control have both been falsified by recent work [84]. Presenting such a claim without evidence is unfalsifiable in the context of that review and therefore constitutes no evidence for or against the report being reviewed.

[83] (Alvino & PLOS ONE Editors, 2018; Baker & Brandon, 2016; Goodman et al., 2016; Kretser et al., 2019; Open Science Collaboration, 2015)
[84] (Ohrens et al., 2019)

Reviewers and editors should also recognize this need for reproducibility and refrain from dismissing scientific communications reporting new tests of already studied effects merely due to lack of novelty (Section III.d).

d. **Independent review and publication bias**

Scientific communications about research are only as valid as the evidence upon which they are based. And the validity of the evidence underpinning scientific communications depends on full disclosure of the methods used to collect the information. As a result, few individuals are capable of vouching for the strength of inference and therefore validity of findings. Although, many individuals review scientific results before they are accepted as evidence. Some of those reviewers are not independent (e.g., co-authors) and others are not qualified (e.g., untrained readers), so a special responsibility is borne by qualified and independent reviewers to follow the principles of scientific integrity we have discussed above.

The scientist(s) who did the research can vouch for all ethical conduct and methods that led to a finding might be the authority for a scientific communication but as the authors they have an intrinsic conflict of interest. Often co-authors on scientific papers (perhaps even senior authors) cannot vouch for every datum, every step of collection and analysis, or each component of scientific integrity behind a given scientific communication. Likewise, independent reviewers (including editors and publishers) who have received privileged information 'behind the scenes' about those methods and findings can, at most, vouch for what has been presented to them and the validity of the inferences drawn from those presentations of methods and data, but not for the entire endeavor. As a result, most if not all scientists, independent reviewers, and certainly third-parties reporting on others' science can only claim partial knowledge. Therefore, no third party, no matter how illustrious, can vouch for a scientific communication.

Even the most strict, thoroughgoing, and formal independent reviewers can typically only speak to the information presented to them. Editors can complement the peer reviewers by oversight of various ethical issues but even editors rely on institutional boards to certify compliance with ethical protocols and conflict of interest disclosures. Third parties that might be able to speak to the ethics and integrity, such as institutional review boards supervising ethical conduct of research or conflicts of interest would be needed to close the loop and vouch for scientific integrity thoroughly. As a result, only a process that involves thorough disclosures by the scientists communicating transparently to institutional officials, peer reviewers, and editors, would be capable of certifying a scientific communication as meeting principles of scientific integrity and standards of evidence. Yet, even such a well-staffed and empowered body might not be immune to external pressures or the possibility of deception by scientists. Therefore, scientific communications depend heavily and ultimately on reproducibility and its own independent review as well as the broader scientific community verifying the claims in scientific communications. That takes time. Independent review is only a gate-keeper therefore and the time-consuming and long-lasting process of advancing knowledge through science cannot be accelerated much without a concomitant loss of reliability.

The incremental growth of knowledge is vulnerable to deception and inadequate independent review. Therefore, ethics and independent review are inseparable. The primary responsibility for independent review falls on the scientists seeking to enter findings as evidence. Clearly professional ethics enter into the compliance with pre-research permissions, compliance with ethical and legal standards throughout the process, and submission of findings to authentic portals for scientific communication. Authentic, independent review as the final step in a scientific communication would depend on third parties (usually editors) choosing by qualified reviewers. We offer a comment on what independent and qualified should mean next.

We define independent as a person who has no shared interest in the success of the scientists submitting the research finding or conflict of interest about the publication or suppression of the evidence. Determining independence can be bewilderingly complex, as testified by the debate now ongoing among scientific journals about how their editors should select reviewers and judge potential conflicts of interest. Adding to the bewilderment, predatory journals, reviewers for hire, and fraud in the review process have begun to interfere with the quality of the scientific literature [85].

Ideally, a qualified expert should have the requisite expertise and no stake in acceptance, revision, or rejection of a scientific communication and its publication. In practice, qualified reviewers often do have a stake, by virtue of belonging to a small community of scientists and siding with one individual or idea over another based on past evidence, personal affiliations, animosities, or ideologies. This paradox (experts might be disqualified reviewers) should lessen as the number of reviewers with expertise rises and their personal relationships become attenuated. The disqualifying aspects of personal preferences may also be attenuated by editorial and publisher practices that anonymize scientific communications or enforce disclosures of conflicts of interest and affidavits of objectivity and the like. Another approach is to seek standards of evidence and disclosures of scientific integrity that are uniform and widely used across many fields, so that qualified experts without conflicts can more easily be found. For example, if the field of wildlife management had embraced randomized, controlled trials decades ago as other fields of science did, many wildlife experiments could be reviewed by experimental scientists from other fields. Such an ideal will never solve the conundrum that particularly specialized research questions might not be amenable to truly independent review. For example, there might be only one expert in the world on the emotional lives of animal z, so any scientific communication touching on that animal’s cognition or experiences might have to be reviewed by that expert or their disciples. In that case, full transparency is recommended, in which authors, reviewers, editors, and publishers are all known to each other and their disclosures of potential conflicts are published. The abductive inference we are making here and in many passages on scientific integrity is inspired by L. F. Brandeis quotation, “Sunlight is said to be the best of disinfectants; electric light the most efficient policeman”.

[85] (Bohannon, 2014; Haug, 2015)

Registered reports: We recommend as much use as is feasible of independent, peer review of scientific methods before the data are collected. These have the advantage of strengthening methods for data collection and analysis before the time and resources are invested. These reviews are referred to as registered reports and a number of journals have offered these tools recently (https://cos.io/rr/). In addition to improving the methods before data collection, registering a study reduces the publication bias [86]. For example, Royal Society Open Science (https://royalsocietypublishing.org/doi/full/10.1098/rsos.160979) – and more recently by a handful of animal research journals to our knowledge (e.g., Biological Conservation and Conservation Biology) – offer a novel approach to registered reports that promotes replication efforts that evaluate the reproducibility of prior findings. The journal does so by a submission policy stating, "Negative findings, meta analyses and studies testing the reproducibility of significant work are encouraged. Experiments with little or no new content will only be considered if they provide a meaningful contribution to the literature, for instance by contributing to reproducibility studies." And by a policy of allowing replication efforts to be submitted as registered reports, while blinding reviewers to whether results are in hand. The hope is that these methods reduce potential reviewer bias for or against confirmatory results, and possible editorial bias against results that do not appear novel, because the "…journal uses Registered Reports as a submission option for replication studies with peer review prior to observing the study outcomes." [87]. We recommend harmful interventions planned in animal research should all be submitted as registered reports, so criteria of scientific integrity and standards of evidence are applied before animals are harmed.

In the majority of cases, most research findings will be evaluated after the fact. For the submitting scientists, a new supplementary form of independent review has emerged in recent years, the pre-print or pre-publication peer review. Pre-prints can be posted online for public scrutiny before, during, and after the traditional, peer review organized by a journal. Such pre-prints are so new that we cannot yet recommend them (see Box 8). At this time, we cannot recommend pre-publication review because we are not aware of methods for ensuring reviewers are qualified, will no spread the preliminary findings publicly, or sway the opinion of authentic reviewers in some way or another.

After submission of research findings to a scientific journal or similar publication, the methods for assuring qualified expert review have also been changing in recent years. Most scientific journals conduct blind, independent review, meaning the authors are unaware of the identity of the reviewers and the editor judges independence of reviewers by journal-specific criteria. (In Sections III.a-c above we described the responsibilities of reviewers under the principles of scientific integrity and noted where an editor might place a restriction on transparency to protect the peer reviewer.). If an independent reviewer might be anyone, one assumes the author will take the review more objectively and thoroughly. In parallel, the reviewer might feel freer to express criticisms if they are protected from retaliation. On the other hand, many of

[86] (Munafò et al., 2017)
[87] (Sanders et al., 2017)

the scientists who contributed to this document have experienced anonymous reviews that are not fair or objective. Anonymity might have been deployed as a shield for unscientific or unprofessional critique. This sentiment seems to be leading a growing number of editors and publishers have begun to use double-blind reviews, in which the identity of the authors is also concealed from the reviewers. In addition, there is some evidence that non-anonymous peer review improves the quality of reviews [88]. We discuss blinding more in Section IV. If disclosures of scientists' conflicts of interest as we discussed previously can be stripped of obvious identifying information, then independent reviewers should not be blind to those when reviewing scientific communications, even if they are made blind to the identity of the scientists. Fundamentally, science should be blind to authority or reputation, and therefore independent review should not either.

Practice in animal research is undergoing gradual change with salutary experiments underway about the process of independent review itself. Therefore, we note that one principle of scientific integrity might be purposefully diminished to enhance another principle. For example, editorial and journal policies that step away from complete transparency can be seen as efforts to improve objectivity of authors and independent reviewers. However, we warn against compromising reproducibility to enhance other aspects of scientific integrity. For example, concealing computer code or other methods to protect confidentiality, intellectual property or top-secret classifications might enhance the objectivity or independence of reviewers, but at the expense of both technical and conceptual reproducibility, which in the long run is the best insurance against weak inference. In sum, we consider experiments with independent review best conducted in the balancing of transparency and objectivity, with one party (usually editors) serving as arbiters who enjoy full disclosure by all parties.

Ultimately, we agree with that trust is required, whether editors trust their reviewers, authors trust their editors, editors trust their publishers, etc. to uphold scientific integrity. Scientists who support integrity and standards of evidence may need to speak with one voice about publishing practices that lower the quality of the scientific literature, such as overloading independent reviewers, fast turn-around times in reviews, and non-transparent favor or disfavor aimed at reviewers by authors and editors [89]. The entire scientific community has an interest in improving practices that enhance the scientific integrity (Section V).

Box 8: Guidelines for independent review in animal research

The only rule we assert is that research findings cannot be considered evidence until they have been subject to independent review by qualified experts.

We do not recommend authors publish research findings before formal, independent review has accepted them as evidence. Likewise, knowing

[88] (Walsh, Rooney, Appleby, & Wilkinson, 2000)
[89] (Alberts, Hanson, & Kelner, 2008; Arns, 2014)

submission of findings to a predatory journal that fabricates independent review or allows cursory review would violate scientific integrity. After traditional, authentic independent review and acceptance by a reputable scientific journal, we do recommend that authors open their work to public scrutiny and comment, consider the commentary carefully, and take appropriate actions to correct, retract, or replicate scientific communications (Section V, Box 11).

Guidelines for independent review: We recommend independent reviewers advocate for double-blind review. Independent reviewers have sway with scientific journals that request pro bono work by reviewers for an ever-increasing volume of scientific communications.

We encourage transparency about editorial policies and editorial scrutiny of independence of reviewers, because of apparent pervasive, undisclosed bias arising from personal rivalries and affiliations between scientists or between practitioners and scientists. We recommend more publishers and editors support registered reports and double-blind reviews.

The methods of independent reviewers are a matter of current debate. Practices are likely to evolve with the addition of new evidence. Therefore, we propose research into a centralized database – such as ORCID https://orcid.org/ or the one for financial disclosures for medical scientists at http://www.convey.org/ – that maintain thorough and up-to-date disclosures of value judgments relating to animals by scientists who wish to serve as reviewers or authors of scientific communications about animal research

IV. Strength of inference and sources of bias

Inference means "the drawing of a conclusion from known or assumed facts or statements; *esp.* in *Logic*, the forming of a conclusion from data or premises, either by inductive or deductive methods; reasoning from something known or assumed to something else which follows from it" [90]. We bracket our discussion here on inferences about cause-and-effect in quantitative animal research with a heavy emphasis on deductive inference from experiments as the most intrusive interventions into animal lives. However, our focus is not only on deductive inference in this document, so we address types of inference in Box 9.

Box 9. Inference: Mainly we address inductive inference "reasoning from particular facts to general principles", but many of our presuppositions and assertions about principles of scientific integrity lead to deductive inference, "reasoning from generals to particulars", as well as admittedly some inference to the most plausible explanation, "the formation or adoption of a plausible but unproven explanation for an observed

[90] (OED, 2018)

phenomenon; a working hypothesis derived from limited evidence and informed conjecture." also called abductive [91]. The three different approaches to inference above are applied in different sections of our text.

Sections I and II apply deductive inference mainly and occasionally abductive inference where we lack evidence or could not find it. But those sections also lay out a roadmap to inductive inference by explaining why we make a recommendation and why it might not always apply. For example, Section III–V are rife with suggestions for scientific practice from our own experience and those of experts; suggestions which might improve standards of evidence but are thus far unproven by inductive inference. Of course, this mix of experiences arises from a mix of inductive and deductive inference itself. Section IV is mainly based on inductive inference because of our focus on quantitative experiments. However, it will be clear that the very approach of suggesting a recipe for improving inference in particular cases is deductive and particularly our encouragement to aim for an as-yet unattained platinum-standard (Section IV.a). In Section V and Box 11 we revert to deductive inference and a heavy dose of abductive inference because relatively little evidence has accumulated so far on correcting the scientific literature [92].

Our conceptual framework for standards of evidence derives from a century of philosophy of science and evaluation of scientific research in many disciplines by numerous authors [93]. We acknowledge the doubts these authors expressed about approximating the truth, but we reject the notion that no scientific evidence can be verified, and the notion that all inferences are equally subjective and mired in epistemic bias. Below, we define standards that increase our confidence in the accuracy of evidence. We begin with strength of inference, explore bias, and propose a single continuum of strength of inference as in Table 1 and Figure 4.

[91] (OED, 2018)
[92] (Marcus & Oransky, 2018)
[93] (Biondi, 2014; Chamberlin 1890; Gawande, 2016; Gould, 1980; Ioannidis, 2005; Kuhn, 1962; Mukherjee, 2010, 2016; Platt, 1964; Popper, 1959)

Table 1. Standard of evidence	**+**	**Platinum**	**+**	**Gold**	**+**	**Silver**	**Bronze**
A. Designs to avoid potential biases→	Blind post-completion	Blind pre-completion	Cross-over design	Randomized	Before-after-remove	Before-and-after comparison	Observation or correlation
Selection		***	*	*			
Treatment		****	***	**	*	*	
Measurement		*	*				
Reporting or Publication	*	*					
B. Substantial shortcomings in scientific integrity or in design can lower the reliability of evidence to silver or below →					No placebo or inappropriate control, Imperfect randomization or blinding		Non-transparency in essential methods, Inconsistent protocols, no reproducibility, no independent review, failure to replicate, failure to correct or retract

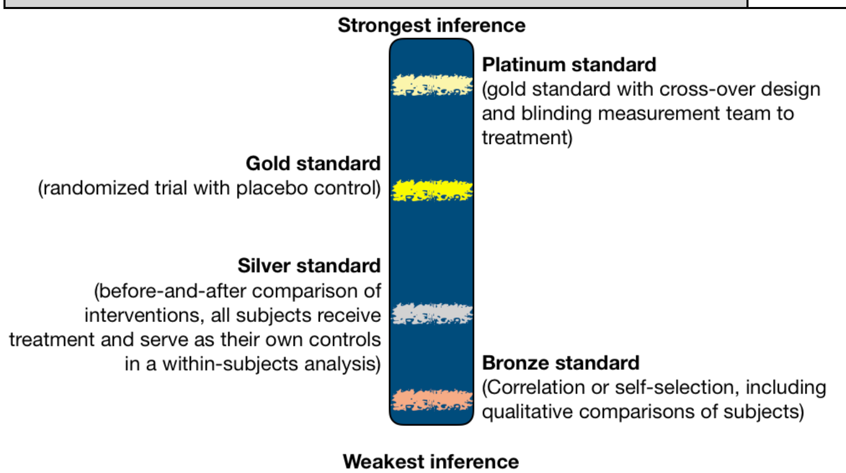

Figure 4. Strength of inference in relation to research design. The positions of standards along a continuum of strength of inference are approximated (fuzzy horizontal bars), because we cannot yet quantify strength of inference. Also, evaluating the strength of inference from a particular study requires close reading of the methods and results. However the relative position of the fuzzy lines for different standards is depicted to reflect the loss of confidence associated with the introduction of confounding variables or the lack of controls, e.g., unbiased, silver standard tests lower the strength of inference by approximately half because all else being equal, they introduce one potentially confounding variable (time). All standards presume no bias sufficient to undermine the reliability of a study.

Our depiction of a continuum of strength of inference (Table 1, Figure 4) in no way undermines our prior assertion that conceptual reproducibility (evaluating cause-and-effect by two different mechanistic pathways, Section III.c) is essential. Exact and technical reproducibility would have the same strength of inference as the original study, all things being equal. Conceptual reproducibility might fall at a different place on the continuum given the scientist has attempted to follow a different route to the same outcome. The depiction in Figure 4 refers to confidence in inferences from a single research effort. A parallel continuum would exist for the independent effort at conceptual reproducibility, in short, a scientist places a given research effort along the continuum by virtue of the design of that effort.

Although we begin with the platinum standard as the strongest inference (Figure 4), we repeatedly refer to elements of the gold standard, because we anticipate that few if any studies in animal research will achieve the platinum standard. The platinum standard includes the gold standard of randomized, controlled trial without bias, but adds cross-over design and blinding as defined below. We treat the platinum standard as an aspirational guideline super-imposed atop the gold-standard, which we deem necessary to strong inference, in almost every case.

Blinding is intended to reduce bias among the scientists and their collaborators [94] . The amount of blinding (single-, double-, triple-, or quadruple-) refers to how many steps in the experiment are concealed from researchers or reviewers. The steps that might be blinded include: (i) those intervening randomly should be unaware of subject histories and attributes and should not communicate which subjects received the control or treatment intervention to others in the research team (this depends on having used an undetectable intervention); (ii) those measuring the effects are unaware of which intervention the subject received (this too depends on having used an undetectable intervention); (iii) those interpreting results are unaware of which subjects received treatment or control; and (iv) those independently reviewing results are unaware of which subjects received treatment or control and unaware of the identity of the scientists who will or have conducted the research. Because science knows no authority, only evidence, blinding independent reviewers to conceal all unnecessary information might avoid several forms of bias (below). Note that blinding steps (ii) and (iii) might be feasibly done by the same set of people but the role in step (i) should be separate from all other roles to assure the success of blinding, and the role in step (iv) should be separate from all other roles to meet the criterion of independence. Experiments with inconspicuous manifestations (e.g., some medicinal treatments) are easiest to blind, but in experiments with long-lasting structural modifications it can be very difficult to conceal conspicuous interventions. For example, lethal methods intended to protect domestic animals from predators are often inconspicuous (e.g., concealed traps) or brief in implementation (e.g., shooting), which would facilitate blinding, whereas many non-lethal methods are conspicuous (e.g., fencing, lights, guardian animals).

[94] Here we broaden our focus to all participants in animal research including owners such as farmers, managers of captive animals, research assistants, etc.

Cross-over design requires the reversal of treatment and control within subjects. Because of randomization, some subjects will begin as placebo controls and others in treatment conditions, but all subjects will reverse to the other condition at approximately the same time midway through the experiment. A third reversal further strengthens inference about the effect of treatment. Therefore, every subject experiences both the intervention and the placebo control. By so doing, excessive differences between subjects are rendered less confounding, by measuring the response of subjects to interventions minus the response of the same subjects to placebo control. Although this might appear to be silver-standard at first glance, it is combined with randomization, so some subjects begin as placebo control and end the study in the intervention group, therefore some subjects experienced change over time followed by treatment, whereas others experienced the reverse [95]. When designing cross-over experiments, it might be important to allow time between the first and reversed treatment for effects to 'wash out' and to account for the possible time lag or long-lasting effects of the treatment. Such 'wash out' periods should be designed at a length appropriate to the effect under study and the memory capabilities of the animal species being affected or replacement time of the individual animals affected.

Randomization is random assignment of subjects to intervention groups or to placebo control groups. Controls are considered essential to making reliable inferences about the effect of an intervention because variability and change are ubiquitous. A placebo control group contains subjects who have received everything but the hypothesized effective treatment and in exactly the same ways, times, and places, e.g., a sugar pill administered just like a medicinal pill, or blank ammunition (i.e., no projectile beyond a few feet from the muzzle of a gun) shot at an animal rather than lethal ammunition. Randomization is widely considered to be the most important step in eliminating bias in experiments because random assignment eliminates the most prevalent and pervasive bias in experiments, referred to as selection bias (below and [96]). Placebo control subjects will change as will those treated, so detecting the effect of intervention against a background of ubiquitous and confounding changes would require a careful comparison of the treated group to the control group.

The exceptions to the gold standard would be rare cases in which a silver-standard is justified by arguments based on ethics, law, or impossibility (Box 10), not convenience or vague references to socio-cultural acceptance. A blatantly unethical or illegal research method might make a gold-standard or better design infeasible. But most arguments about feasibility should pass the test we laid out for authentic impossibility. We detail the guidelines for claiming a gold-standard design is impossible after describing the standards fully below.

Silver standard is defined as before-and-after comparisons of interventions. In silver standard studies, every subject gets the intervention (no placebo control) and each subject is compared to itself before intervention. For example, the number of domestic animals lost prior to

[95] (Ohrens et al., 2019)
[96] (Ioannidis, 2005)

intervening is subtracted from the number of domestic animals lost after intervention. Before-and-after comparisons are also called case-control experiments or BACI (before and after comparison of impacts without randomization) and are often used when randomizations or placebo controls are considered infeasible (Box 10).

Before-and-after comparisons are also called case-control experiments or BACI (before and after comparison of impacts) and are often used when randomization is considered infeasible. If BACI includes randomization, we refer the reader to the gold-standard above. Much has been written on stronger and weaker inference in BACI designs (Murtaugh, 2002; Stewart-Oaten, 2003), with a good example in a related field to ours (Popescu, de Valpine, Tempel, & Peery, 2012). Statisticians seem to us to have reached consensus that non-random BACI designs should employ first-order (at least) serial autocorrelation statistics which treat within-subject measurements as time series and consider expert information on local events that might confound effects of treatment and the proportion of subjects so affected relative to total sample size.

Silver is a lower standard than gold because inference is weaker. At a minimum, silver-standard studies introduce a new variable, time, i.e., all subjects underwent the passage of time that affects individuals differently. Consider the analogy of a cold remedy. We know most people recover from colds over time. Therefore, any proposed treatment should work faster or better than the natural, healthy person's recovery from a cold. If the putative treatment for colds is tested by a silver-standard design, the inference that it was effective is difficult to distinguish from the inference that subject patients got better on their own as time passed. Non-randomized BACI might have difficulty distinguishing treatment effect from time effect if selection bias was introduced in who received the cold remedy (e.g., patients who volunteer for an experimental remedy are usually not a random sample of patients (Mukherjee, 2010)). Predator control experiments are often good analogies to the hypothetical cold remedy. Domestic animals might be attacked by predators only once with no repeat, even in the absence of intervention (see previous section on uncertainties in predator control). Therefore, loss of a domestic animal might not be repeated simply because of the passage of time. The uncontrolled effect of time passing is why we rate silver-standard designs as producing inference that is half as strong as gold-standard designs. The presence of a control, comparison group chosen without selection bias is therefore essential to raising the strength of inference.

Predator control experiments are good analogies to the hypothetical cold remedy above. Domestic animals might be attacked by wild or feral predators only once with no repeat, even in the absence of intervention. Observations of such 'no-repeat' predation events have been made since 1983 if not earlier [97]. Therefore, loss of a domestic animal might not be repeated simply because of the passage of time, regardless of any intervention. The uncontrolled effect

[97] (Tompa, 1983)

of time passing is why we consider silver-standard designs produce inference that is half as strong as gold-standard designs.

It is worth acknowledging that one can improve on silver-standard somewhat if one staggers treatment so that subjects do not all experience treatment at the same time. Such staggering might eliminate the simultaneous effect of a brief confounding effect on all subjects. Nevertheless, subjects still experience time passing even if not simultaneously. Researchers have addressed the confounding effect of time passing by removing treatment and monitoring their subjects again so there are three measurements at least: before-treatment baseline, after-treatment response, and after removal of treatment another response. While stronger than before-and-after comparisons, we still see two problems with recommending this design: First, the ability to remove treatment in the final phase implies the researcher has influence to manipulate the treatment, which begs the question why not treat randomly? Perhaps, the treatment is not under the influence of the researcher, but it ends for all subjects simultaneously or after a predetermined duration. If so, we would refer to this as 'silver+'. Yet the slightly stronger inference of silver+ merits scrutiny for a second reason. The variable 'time' still affected every subject in parallel with the treatment, so we have an n =2 for the effect of time. If one wants strong inference about the effect of time independent of treatment one needs a higher n of re-treatments and removals. That would seem to drag out the trials and once again beg the question of why not work harder to randomize and cross-over? Therefore, we conclude that before-during-after does not improve much on silver standards, only approximating gold standard with a large number of treatments and removals.

To generate reliable inference from the topmost or lower standards of platinum-, gold-, or silver-standard experimental design, a scientist should also eliminate four types of bias [98]: In the following examples, we rely heavily on the literature about interventions to protect domestic animals from predators (hereafter predator control):

a. **Selection bias (also known as sampling bias).** To avoid this bias, subjects should be assigned to treatment and placebo control randomly from the same set of candidate subjects. Selection bias can arise when the choice of which subject receives the treatment and which subject receives the placebo control is non-random (or when the sample is so small that even randomization cannot prevent intervention groups from differing significantly at the outset). Selection bias is rife in predator control research because domestic animal herds are often selected by the owners or by experimenters to receive a treatment or not. Selection rather than randomization undermines strong inference about an intervention effect because subjects naturally vary in their response to an intervention and the circumstances surrounding them might influence the effects of a treatment. Thus. selection bias might lead to a treatment group that is more likely to respond in the predicted way to the intervention. Self-selection is a form of selection bias that has long been recognized as slanting results severely in fields as distinct as

[98] (Ioannidis, 2005; Treves et al., 2016)

sociology and medicine. But experimenters have also been implicated in selection bias when they intentionally or unintentionally assign subjects non-randomly [99]. For example, experimenters treating valleys with aerial gunning of coyotes attempted the treatment in all valleys but could not do so in all valleys for reasons that were not clearly explained. Post hoc the researchers labeled those valleys as controls [100], but such selection by the researchers is inherently biased because the inability to treat the valleys for whatever reason was the sole criterion for selection of the control group. Indeed, that study epitomizes an array of biases useful for students of experimental design [101].

b. **Treatment bias.** This occurs where the intervention or placebo controls are administered haphazardly. It can be avoided by strict quality controls. A common form of treatment bias in predator control is to tailor the intervention method or its intensity to the subjective impressions of the domestic animal owners or the agents implementing. For example, even the best experimental tests of lethal methods for predator control fail to distinguish the techniques applied, e.g., pooling shooting, trapping, poison, poaching, or regulated hunting, into one category of intervention. If care in standardizing interventions is not taken, it is easy for implementers to put more effort into subjects that seem to need more, or distribute the intervention by convenience, both of which are likely to produce systematic error [102].

   Treatment bias (and additional limits to inference) might increase when placebo controls are not very similar to interventions: For example, a medicinal treatment might be detectable to the subject from side-effects while the placebo control might produce no side-effects. Or, shooting with lethal ammunition might not be adequately mimicked by shooting with placebos of sub-lethal ammunition or blanks, if the placebos do not have the same range or sound and light effects of live ammunition. The result of inadequate placebo controls might be lower confidence in isolating the cause of a particular effect (hence silver or silver+ standards in Table 1). For example, if the medicinal treatment causes nausea as a side-effect, but the placebo control does not, might any health benefits of the treatment instead relate to the body's endogenous response to nausea rather than the exogenous effects of the medicine? We also address placebos under measurement bias next.

c. **Measurement bias.** To avoid this bias, one should ensure that measurements are taken uniformly across intervention groups and placebo control groups. Ideally, those

[99] (Ioannidis, 2005; Mukherjee, 2010; Treves et al., 2016)
[100] (USDA-WS & others, 2016; Wagner & Conover, 1999)
[101] (Treves et al., 2016; "Western Watersheds Project et al. v USDA Wildlife Services," 2018)
[102] (Ioannidis, 2005; Mukherjee, 2010; Treves et al., 2016)

collecting data on the intervention and the placebo control are unaware of which the subject received (blinding).

Measurement bias might also arise if the ‘placebo effect’ is strong and subjective measures of effect are used. In brief, the placebo effect is the widespread phenomenon that human subjects often respond positively to participating in an experimental trial, regardless of whether they receive an intervention or a placebo. The placebo effect can produce conservative error when the subjects receiving the placebo control feel better, so the apparent effect of the intervention is slight. But the placebo effect can inflate the apparent effect of intervention if the placebo control is not adequately simulating the intervention. When owners of animals self-report effects (subjective) the risk of placebo effects would seem to rise more than in designs with objective measures of effects.

d. **Reporting bias**. If scientists report results unevenly, omit data or methods, or report in a way that is not even-handed with regard to intervention or placebo control, one gets reporting bias. Reporting bias is most easily avoided if the analysts and reporters of a study are unaware of whether the subject received intervention or the placebo control until after the aggregate statistics are analyzed. It also helps to have more co-authors and peer reviewers willing to review the data, analyses, and interpretations, as well as to pre-register a study, which should generally reduce publication bias against negative results.

   Also reporting bias should be lessened if scientists understand and transparently report the limits to inference inherent to the methods and research design used in a particular research effort. For the purposes of animal research, a basic issue is that the results of a given data collection effort are insuperably limited to the methods used, possibly including the time, subjects, places, and materials, and perhaps even the observers involved (Box 5). Furthermore, research on wild populations of animals is inherently challenging since many factors are dynamic, cannot be controlled for, nor are they fully understood in many cases. Limits to inference should be signaled in scientific communications, e.g., we infer that x caused y in our experimental conditions and predict that manipulating x would change y in other circumstances. Readers will recognize the semantic over-generalization of the alternative (e.g., x caused y so modifying x will alter y). It is objectivity that makes scientists think cautiously about their research, but we present this text under reporting bias because successes and failures arise during scientific communication. We expect that the problem of exceeding the limits to inference are a common result when study designs do not have adequate comparison subjects or controls are not sufficiently similar to interventions to act as a placebo. We return to the limits to inference below, in the context of placebo controls and strong inference.

e. **Publication bias.** Because publication bias ,when it occurs, is expressed by editors or independent reviewers, it can only be influenced by selection of an appropriate outlet for scientific communications by the scientists designing the research. Therefore, we discuss publication bias in Section III.d and Section V.

If a study uses randomized design, cross-over, and blinding, and avoids all four biases above, it meets the platinum-standard. If a study both uses a randomized design and avoids the four biases, it satisfies the gold standard. Gold-standard experiments present somewhat weaker inference than platinum-standard experiments because the cross-over design guards against chance differences in the intervention and placebo control groups. Crossover design is particularly important when the samples of subjects are small. Also, blinding can reduce all forms of bias to the extent they are extended over the appropriate parties. In Table 1 explicitly and in Figure 4 implicitly, we allow for 'silver+' and 'gold+' designs that include some but not all of the elements of the platinum standard. To our knowledge, scholars of experimental design have not yet developed a method to quantify the gain or loss of inference associated with different designs, therefore we impose a qualitative system of increment or decrement to guide those planning or evaluating research designs.

In wild animal research, we have encountered several common arguments for not designing experiments using gold-standards without bias. The most common is that inherent variability of subjects is too great. Essentially, this argument is that any intervention effect would be less than the pre-existing differences between subjects, even when randomly chosen. *Prima facie*, this argument is weak because it suggests the intervention will have a small effect (i.e., it is not a very effective intervention). But the argument also fails on the grounds of being unfalsifiable as a general claim (it does not specify which subjects are too variable or which response variables have too high variance). In sum, this argument needs to be stated precisely before it can be falsified and probably it can only be falsified by trying a randomized, gold-standard experiment. Moreover, the biomedical sciences have long understood that one can overcome situations of high between-subject variability by adopting a cross-over design or increasing the sample size [103]. The second most common objection we have encountered is that subjects (or their owners in the case of animals) will refuse the placebo control. Again, this is difficult to falsify without a transparent and conscientious effort. We reject it for our present purposes because animals will not reject the placebo control, probably by definition wishing to avoid the intervention in the first place. Owners of animals might exert proprietary rights to reject a placebo, but experimenters are urged to use persuasion to explain why it is necessary as we have done above. Nevertheless, we identify a few situations below in which the gold-standard might indeed be impossible.

Box 10: The feasibility of achieving the strongest inference

Given the variety of situations in which animal research might be conducted, it is conceivable that a research team would find it impossible to design a platinum-

[103] (Díaz-Uriarte, 2002; Ioannidis, 2005)

standard experiment, a gold-standard experiment, or eliminate all potential biases.

To provide guidelines for situations in which gold- or platinum-standard experimental designs might be deferred until feasible, we have to differentiate feasibility from impossibility.

Feasible ("Of a design, project, etc.: Capable of being done, accomplished or carried out;…." [104]) should not be confused with impossible ("Not possible; that cannot be done or effected; that cannot exist or come into being; that cannot be, in existing or specified circumstances." [105]).

The common usage of impossible reflects a person's perception that they do not have the capability, time, or resources to accomplish something, in addition to authentic impossibility. We aim to separate those concepts as much as possible here, because animal research almost always involves an infringement of that organism's interests, so we should be aware when a human claim something is impossible when it is actually infeasible for the human.

Authentic impossibility means one of two things: (1) that two actions or events are mutually exclusive although either is feasible (e.g., I cannot study the behavior of an animal and study its death at the same time); or (2) an action or event would violate physical laws (e.g., I cannot survive in outer space without a space suit). The latter example acknowledges that some technological innovations and scientific discoveries overcome former impossibilities, which underscores the distinction between 'action x is impossible' and 'action x is not currently feasible'. For practical purposes, most people's response to difficult situations can be rephrased as 'I do not currently have the motivation, legal authority, time, skills, or resources to accomplish that action'. That is not the same as impossible because the obstacles might change in short order.

Although many actions are authentically impossible, most objections to improving the standards for animal research are actually claims of feasibility. Few elements of the platinum standard or gold standard without bias are impossible. Rather they can be very difficult, and difficulty might make such designs infeasible but claims of infeasibility demand scrutiny.

We present a short non-exhaustive list of research design elements that pertain to platinum- and gold-standard experiments, which might be authentically impossible in

[104] (OED, 2018)
[105] (OED, 2018)

particular circumstances. We see the following elements of platinum-standard and gold-standard designs as authentically impossible:

- Cross-over design in which subjects experience lasting effects of manipulation (e.g., injured, physically modified, or killed) cannot be considered placebo controls after the intervention until the effects of modification are confirmed to have passed (see our discussion of alternatives to lethal animal control in Box 5). We recommend longer periods of 'washout' or recovery after interventions that are more invasive.

- Blinding research teams when interventions are conspicuous, long-lasting, or subjects are permanently changed by intervention appears to nullify blinding. We recommend other stringent approaches to eliminate measurement bias, such as interobserver reliability or unannounced inspections by third-party verifiers.

- Sociocultural conditions might render specific elements of platinum-standard designs impossible. However, we caution that most sociocultural constraints are issues of feasibility, i.e., one fears that efforts at persuading people will backfire if seen as coercive.

- Biophysical conditions deserve attention as they might be immutable. For example, some intervention subjects may necessarily spend time in close proximity to placebo control subjects and thereby the intervention and placebo control might 'bleed' into each other. Although this produces treatment bias, it might be tolerated without losing the gold-standard because it can generate conservative error (the intervention effect resembles the placebo control effect because subjects from the intervention and placebo control groups interacted). When the treatment bias is not conservative, (i.e., differences in effect for treated subjects and placebo controls are inflated), the bias is so damaging to inference that one cannot hope for better than silver-standard design (Table 1). Therefore, when experimenters deem it impossible to segregate animal subjects into independent intervention groups and placebo control groups, we recommend selecting different sites where such segregation is possible or settling for silver-standard design in which all subjects receive the intervention after baseline measures of their "control" conditions.

The bulleted items above are not meant as an exhaustive list but rather as examples of situations in which platinum- or gold-standard will be authentically impossible. In these cases, one might accept a lower standard of inference as evidence, until technological or scientific discoveries make the higher standard possible. The above reasoning is distinct from the assertion that experimenters cannot feasibly use cross-over design, blinding, random assignment, or purge bias. Most such arguments place the proposed research in a vulnerable position scientifically. While we recognize that partnerships, costs, time, expertise, materials, and sufficient samples of subjects are all limiting factors in research design, these are issues of feasibility that suggest the need for greater investment or alternative pathways to the same information. Failure to find alternatives or presenting weaker inference as evidence has ethical and legal implications.

Finally, we have not addressed in detail the weaker inference of correlation (bronze-standard) or the many forms of observational and qualitative research. We do not dismiss these necessary initial modes of scientific investigation. Rather, we point out that they are useful in developing and opposing reasonable, alternative hypotheses, or identifying gaps in knowledge that demand further data collection. But we have doubts that science without careful comparison of treatment to placebo controls can produce strong inference in animal research, such as that needed to justify human actions that further intrude on nonhuman lives.

Box 10: Guidelines for strong inference in animal research

Animal research should use randomization, control, and avoid biases. We recommend the platinum standard experiments without bias, as a guideline. If one of the 2 elements of the platinum standard that improve it over the gold standard is infeasible (defined above), we recommend the other element be used, as a type of gold+ standard (Table 1, Figure 4).

As a rule, gold-standard experiments without bias are the required strength of inference. However, we recognize situations might preclude gold-standard experiments. Above, we provide guidelines for scientists and independent reviewers to evaluate if authentic impossibility precludes a gold-standard experiment. When silver-standard experiments are conducted, it should be clearly explained why platinum- or gold-standard was infeasible to reach and consider lower standards of inference in the interpretation of results and presenting conclusions. For all work, scientists should assiduously avoid bias. Even the best platinum-standard design is vulnerable to bias in selection, treatment, measurement, or reporting. Publication bias is a real hazard for reliable, scientific communications (Section III).

Guidelines for independent review: Reviewers of proposed research and completed research submitted for publication should be aware of the distinctions between feasibility and impossibility and ask the probing questions that will expose which is which. They should also take the strength of inference into account when evaluating interpretation and conclusions by the authors.

> Reviewers should not feel compelled to approve research simply because it is novel or communicated by authorities. Each such communication is a stand-alone effort judged by the same criteria as all other communications.
>
> Independent reviewers should be aware of the additional pitfalls to strong inference identified by Ioannidis [106] in randomized trials in biomedicine. He pointed out that intervention effects with small effect sizes and little power are prone to positive detection bias (an intervention effect is inferred when none exists), when research is attractive to media or to donors, when results have potential widespread and profound consequences, or when the scientist's career depends on publishing in high-ranked journals.

V. Correction, retraction, intimidation, and retaliation

The scientific literature does not clean itself. Although science progresses, it does not delete studies with weak inference or studies with fabrication, falsification, plagiarism or other acts of intentional and unintentional research misconduct. This duty falls to the scientific community.

The task of cleaning up weak inference and unethical science is urgent because animal research, and the interventions that stem from it, infringe on the lives and well-being of animals. The task is also important for that reason, and because current humans owe future humanity an unimpaired legacy of nature. Owners of captive animals might have a similar interest in solid scientific evidence to guide ethical action towards their animals. Furthermore, public scientists are accountable for transparent, sophisticated accounting to the broadest public, which includes correcting their own research findings and scientific communications [107].

We do not mean to suggest in this section that the scientific community should seek to retract weak inference once replaced by stronger inference. That would obliterate the history of scientific progress and could lead to deeper problems if the new ostensibly stronger evidence proved unreliable. But the users of scientific research, such as government agencies, managers, and the public, should be aware that not all scientific findings are equal. Therefore, one should not list weak inference alongside strong inference as equally valid. Some readers may view this as not going far enough because failing to purge the literature might lead secondary sources to communicate weak inference. But we reserve correction and retraction for serious flaws as described below.

Researchers have primary responsibility for correcting, retracting, or publicly expressing diminished confidence in their own scientific communications, no matter how old the findings might be. An interesting, new test of this idea is the loss of confidence project [108]. The latter

[106] (Ioannidis, 2005)
[107] (Sax, 1970; Treves et al., 2017)
[108] https://lossofconfidence.com/

project offers a space for authors to share a loss of confidence in their own work, without necessarily going to the time-consuming and stigmatized path of submitting corrections to a journal. Apart from cleaning the literature, one goal for such a project is to reduce the stigma associated with corrections and retractions of published papers and thereby motivate more scientists to clean up their own errors. Of course, it is hard for a scientist to identify flaws or publicize corrections or a loss of confidence in their own research. Therefore, scientists should welcome review by qualified experts, not side-step such review by omitting citation to such experts or by explicitly discouraging those experts as reviewers or replicators. Likewise, conscientious responses to reviewers, critics, and replication efforts is an ancillary duty for scientists conducting animal research.

Therefore, the scientific community, qualified reviewers, the editors, and the publishers play important, secondary roles in encouraging self-correction. The broader scientific community has secondary responsibility for cleaning the scientific record if credible evidence surfaces of omissions, errors, or other misconduct (fabrication, falsification, or plagiarism). This topic has recently grown in prominence with the formation of Retraction Watch, an NGO, and wide subscription to the Committee on Publication Ethics and its attendant guidelines (COPE 2019 https://publicationethics.org/about/our-organisation accessed 1 June 2019). Indeed, reviewers, editors, and publishers have specific responsibilities for the quality of scientific communications. However, in our experience, still too many editors and publishers in animal research pay lip service to addressing publication ethics and the many concerns about lack of transparency, objectivity, reproducibility or authentic independent review. Therefore, we expect that post-publication review by qualified experts will be necessary for some time to come, as a long-lasting safeguard against unreliable evidence.

Sometimes scientists challenge each other's findings in public communications. Such challenges traditionally played out in the pages of journals. But challenging work on the basis of bad faith or an appearance of research misconduct might be difficult to publish, might not achieve the desired goal if a culprit is given a chance to rebut, and sometimes requires confidential investigation of unpublished material by a journal's editors or the author's institution. Also, the volume of articles published annually makes it infeasible to clean the scientific literature by submitting challenges for peer review. Therefore, some skeptical, independent reviewers who doubt the veracity of scientific communications might elect to petition editors for correction or retraction. Evidence as not yet been collected on whether research misconduct is better addressed through public communication's first or by private communication's with editors, original authors, publishers, etc.

Bad faith by scientists, independent reviewers, editors, or journal publishers would occur if the appearance of fabrication, falsification, plagiarism or other misconduct were not addressed transparently, swiftly, and thoroughly by exposing and correcting the mis-steps. Likewise, efforts to silence criticus or retaliate against them, withhold data from them, or disparage the critics should all be viewed as potential misconduct and investigated accordingly. For such cases, we recommend whistle-blowing.

Whistle-blowing is alerting any of the appropriate authorities (editors, publishers, institutional oversight committees, government regulators, etc.) when breaches of scientific integrity are suspected. Whistle-blowers and critics in general take a risk in sharing their concerns about scientific integrity or research misconduct. There has been increasing attention to efforts to silence critics or ignore qualified criticisms. Intimidation, silencing and retaliation against critics by any party are unacceptable and suggest that the research being protected should be suspected. Yet, aggressive individuals or organizations who criticize research may be mistaken for trying to intimidate or retaliate against rivals. Conversely, efforts by any individual or organization to stifle criticism or scientific debate under the guise of maintaining harmony or team cohesion should be regarded as silencing academic freedom and open debate, which are hallmarks of science done with integrity [109]. he litmus test in our view for constructive criticism versus intimidation and retaliation is if the methods, data, results, or interpretations of scientific communications are being questioned or criticized versus the individual identities, reputations, motives, funding, associations, institutions, or characters of the scientists are being questioned or criticized. Because that can be a fine line, allegations about perceive intimidation or retaliation against critics should lead oversight organizations to investigate possible misconduct by the retaliators. We direct readers to the National Academies 2017 recommendations on fostering scientific integrity, to identify intimidation or retaliation, and appropriate responses by appropriate oversight organizations (https://www.nap.edu/catalog/21896/fostering-integrity-in-research accessed 1 June 2019).

Box 11. Correction, retraction, and sanction

We propose it is a duty to clean the literature on animal research of unethical science, distinguish weak inference from strong, and repudiate scientific communications based on misconduct. How a scientist or other entity does so, is a matter for guidelines not rules as follow.

Scientists should clean their own scientific communication about evidence by publishing corrections or retractions of their own work. This would seem appropriate when there are no errors, but good faith findings are no longer trusted by the scientists conducting the research. I

For third parties, if one thinks an error was made in good faith, we recommend first contacting the authors directly in hopes they will correct r publicize a lack of confidence. If open dialogue fails to resolve differences, then we recommend other avenues. Distinguishing errors in good faith from poor scientific integrity is not easy and demands a comprehensive look at the process of research that was followed. For example, transparently publishing all data and later finding an error in those data that leads to an error in interpretation was most likely done in good faith because the transparency facilitates the later discovery of the error

[109] (Ioannidis, 2018)

rather than concealment of data at the outset. An error in which data are withheld without good reason or requests for the data are refused or delayed for unreasonably long periods, suggest bad faith and undermine reliability of the original findings. Also, long delays to correct demonstrated errors or to seek harmony through private agreement between critics and authors rather than correction, both seem to approach bad faith.

Evidence of deliberate misconduct (falsification, fabrication, and plagiarism as defined by the U.S. False Claims Act [110] and other U.S. federal laws) define illegal violations of scientific integrity. Short of illegal violations of scientific integrity, gross violations might have implications for boards of ethics or institutional committees overseeing research ethics. Whistle-blowers face risks for their courage and vigilance, so we recommend the highest level of confidentiality be granted them, while appropriate authorities investigate. National Academies 2017 recommendations on fostering scientific integrity provide useful guidance on research misconduct and whistle-blowing https://www.nap.edu/catalog/21896/fostering-integrity-in-research.

Special duties of public scientists: We define public scientists as those whose education or research was funded by the public or by government at any time in their career. At present, this duty is only a professional, ethical duty in our view. However, public science might one day entail a trust duty that is enforceable under U.S. law, following the U.S. Supreme Court in Illinois Central 1896 [111] and the U.S. Uniform Code of Trusts [112]. We set aside the legal obligations and suggest the guideline that public scientists are trustees of the broadest public (current and future generations of the nations in which they work or live). Accordingly, the trustee's duty is to report scientific evidence and review scientific communications with the highest possible standards of evidence and scientific integrity. With that comes an obligation to replace weak inference with strong, and purge errors and the products of research misconduct, so the scientific literature continually improves as a source of strong evidence for public policy discussions.

Guidelines for independent review: Independent reviewers like all scientists have a duty to clean the scientific literature. The confidentiality imposed by journals in the anonymous peer review context might lead to a limited whistle-blowing in which independent reviewers alert editors and publishers to apparent breaches of scientific integrity. Such actions become imperative if there is an appearance of gross violations rising to scientific misconduct. Editors and

[110] https://en.wikipedia.org/wiki/False_Claims_Act
[111] ("Illinois Central Railroad Company v State of Illinois," 1892)
[112] For a summary see http://uniformlaws.org/ActSummary.aspx?title=Trust%20Code

publishers have a more stringent duty to investigate those allegations with due diligence and set aside confidentiality once the investigation is complete, in favor of the highest standards of evidence for the broadest public and decision-makers, as per COPE guidelines 2019 https://publicationethics.org/about/our-organisation accessed 1 June 2019).

VI. Conclusions

Throughout this document, we laid out our interpretations of scientific integrity and strength of inference, including 11 Boxes of specific rules and recommended guidelines. In summary, we feel the combination of scientific integrity and strength of inference together establish standards of evidence for animal research. Although we have not yet articulated fully a logical method to discriminate two pieces of evidence that both meet the minimum standard, we do have confidence that this document captures that minimum without undue subjectivity. To wit, we feel that any evidence from animal research that conforms to the four principles of scientific integrity (transparency, objectivity, reproducibility, and independent review in that order of priority) and also meets our criteria for strong inference: preferably gold-standard design without bias, or failing that, with an adequate justification for silver standard design without bias. Neither scientific integrity alone nor strong inference alone will suffice. Deficiencies in one or the other might not be disqualifying. We acknowledge the remaining elements of subjectivity to our criteria (Table 1, Figure 4), which remain to be determined by independent reviewers in the short-term and by the scientific community and others in the longer-term.

VII. References

1. Ajzen, I. (1991) "The theory of planned behavior," *Organizational Behavior and Human Decision Processes,* 50: 179-211.
2. Alberts, B., Hanson, B., & Kelner, K. (2008) "Reviewing peer review," *Science,* 321: 15.
3. Alvino, G., & PLOS ONE Editors. (2018) "Building research evidence towards reproducibility of animal research," In *PLOS ONE blog News & Policy*. https://blogs.plos.org/blog/2018/08/06/building-research-evidence-towards-reproducibility-of-animal-research/,
4. Arns, M. (2014) "Open access is tiring out peer reviewers," *Nature,* ;515: 467.
5. Artelle, K. A., Reynolds, J. D., A., T., Walsh, J. C., C., P. P., & Darimont, C. T. (2018) "Hallmarks of science missing from North American wildlife management.," *Science Advances* 4: eaao0167.
6. Baker, M., & Brandon, K. (2016) "1,500 scientists lift the lid on reproducibility," *Nature,* 533: 452-454.
7. Batcheller, G. R., Bambery, M. C., Bies, L., Decker, T., Dyke, S., Guynn, D., et al. (2010) "The Public Trust Doctrine: Implications for Wildlife Management and Conservation in the United States and Canada," In. Bethesda, MD: The Wildlife Society, the Association of Fish and Wildlife Agencies (AFWA), the Western Association of Fish and Wildlife Agencies (WAFWA), and the Wildlife Management Institute (WMI).

8. Bekoff, M. (1996) "Cognitive ethology, vigilance, information gathering and representation: Who might know what and why?," *Behavioural Processes,* 35: 225-237.
9. Berger, K. M. (2006) "Carnivore-livestock conflicts: Effects of subsidized predator control and economic correlates on the sheep industry " *Conservation Biology,* 20: 751-761.
10. Bernstein, R. J. (1991) *Beyond Objectivism and Relativism: Science, Hermeneutics and Praxis*. Philadelphia: University of Pennsylvania Press.
11. Biondi, F. (2014) "Paleoecology grand challenge," *Frontiers in Ecology and Evolution,* doi: 10.3389/fevo.2014.00050.
12. Blumm, M. C., & Guthrie, R. D. (2012) "Internationalizing the Public Trust Doctrine: Natural Law and Constitutional and Statutory Approaches to Fulfilling the Saxion Vision," *UC Davis Law Review,* 45: 741-808.
13. Blumm, M. C., & Paulsen, A. (2013) "The Public Trust in Wildlife," In *ExpressO Available at: http://works.bepress.com/michael_blumm/16*.
14. Blumm, M. C., Wisehart, L., Stein, E., Brown, I., Guthrie, R. D., Smith, M., et al. (2014) "The Public Trust Doctrine in Forty-Five States (March 18, 2014)." *Lewis & Clark Law School Legal Studies Research Paper*.
15. Blumm, M. C., & Wood, M. C. (2017) ""No Ordinary Lawsuit": Climate Change, Due Process, and the Public Trust Doctrine," *American University Law Review,* 67: 1-83.
16. Bohannon, J. (2014) "Who's Afraid of Peer Review?," *Science,* 342: 60-65.
17. Carroll, C., Hartl, B., Goldman, G. T., Rohlf, D. J., Treves, A., Kerr, J. T., et al. (2017) "Defending scientific integrity in conservation policy processes: Lessons from Canada, Australia, and the United States," *Conservation Biology,* 31: 967–975.
18. Chamberlin , T. C. (1890) "The method of multiple working hypotheses," *Science,* 15 (reprinted in 1965, vol. 148): 92–96 (reprinted p. 754-759).
19. Clark, S. G., & Milloy, C. (2014) "The North American Model of Wildlife Conservation: An Analysis of Challenges and Adaptive Options," In *Large Carnivore Conservation: Integrating Science and Policy in the North American West,* edited by S. G. Clark & M. B. Rutherford. Chicago: The University of Chicago Press.
20. Darwin, C. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.
21. Díaz-Uriarte, R. (2002) "Incorrect analysis of crossover trials in animal behaviour research," *Animal Behaviour,* 63: 815-822.
22. Durham, W. (1991) *Coevolution: Genes, Culture and Human Diversity*: Stanford University Press, Stanford.
23. Eklund, A., López-Bao, J. V., Tourani, M., Chapron, G., & Frank, J. (2017) "Limited evidence on the effectiveness of interventions to reduce livestock predation by large carnivores," *Scientific Reports,* 7: 2097 | DOI:2010.1038/s41598-41017-02323-w.
24. Fishbein, M., & Manfredo, M. J. (1992) "A theory of behavior change," In *Influencing human behavior: Theory and applications in recreation, tourism, and natural resources management,* edited by M. J. Manfredo. Champaign, Illinois: : Sagamore Publishing, Inc.
25. Author (2016) "The Mistrust of Science," *New Yorker* 10 JUNE 2016.
26. Geer v State of Connecticut. 1896. U.S. Supreme Court. 161 U.S. 519. 161 U.S. 519.

27. Geist, V., Mahoney, S., & Organ, J. F. (2001) "Why hunting has defined the North American model of wildlife conservation," *Transactions of the North American Wildlife and Natural Resources Conference,* 66: 175-185.
28. Goodman, S., Fanelli, D., & Ioannidis, J. (2016) "What does research reproducibility mean?," *Science Translational Medicine,* 8: 341ps312 DOI: 310.1126/scitranslmed.aaf5027.
29. Gould, S. J. (1980) "The promise of paleobiology as a nomothetic, evolutionary discipline," *Paleobiology* 6: 96-118.
30. Hair, K. H., MacLeod, M. M., & Sena, E. (2018) "A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus)," *PloS One blog*.
31. Hare, B., & Tomasello, M. (2005) "Human-like social skills in dogs?," *Trends in Cognitive Sciences,* 9: 439-444.
32. Haug, C. J. (2015) "Peer-Review Fraud — Hacking the Scientific Publication Process," *New England Journal of Medicine,* 373: 2393-2395.
33. Howard, W. E. (1988) "Why lions need to be hunted," *Proceedings of the Mountain Lion Workshop,* 3: 66-68.
34. Hughes v Oklahoma. 1979. U.S. Supreme Court. 441 U.S. 322.
35. Illinois Central Railroad Company v State of Illinois. 1892. U.S. Supreme Court. 146 U.S. 387.
36. Ioannidis, J. P. (2005) "Why Most Published Research Findings Are False," *PLOS Medicine,* 2: e124.
37. Ioannidis, J. P. (2018) "Cochrane crisis: Secrecy, intolerance and evidence-based values," *European Journal of Clinical Investigation*: e13058.
38. Kinzig, A. P., Ehrlich, P. R., Alston, L. J., Arrow, K., Barrett, S., Buchman, T. G., et al. (2013) "Social Norms and Global Environmental Challenges: The Complex Interaction of Behaviors, Values, and Policy," *Bioscience,* 63: 164–175.
39. Klein, e. a. (preprint 2018) "Many Labs 2: Investigating Variation in Replicability Across Sample and Setting," *PsyArXiv* https://psyarxiv.com/9654g/
40. Kretser, A., Delia Murphy, Stefano Bertuzzi, Todd Abraham, David B. Allison, Kathryn J. Boor, et al. (2019) "Scientific Integrity Principles and Best Practices: Recommendations from a Scientific Integrity Consortium," *Science and Engineering Ethics*: 1–29.
41. Kricher, J. (2009) *The Balance of Nature: Ecology's Enduring Myth*. Princeton, NJ: Princeton University Press.
42. Kuhn, T. S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
43. Lennox, R. J., Gallagher, A. J., Ritchie, E. G., & Cooke, S. J. (2018) "Evaluating the efficacy of predator removal in a conflict-prone world," *Biological Conservation,* 224: 277-289.
44. Lynn, W. S. (2004) "The Quality of Ethics: Moral Causation in the Interdisciplinary Science of Geography," In *Geographies and Moralities: International Perspectives on Justice, Development and Place* edited by T. I. R. Lee & D. M. h. Smit. London: Routledge.
45. Lynn, W. S. (2015) "Being Animal," *Society & Animals* 23: 421-424.
46. Lynn, W. S. (2018) "Bringing Ethics to Wild Lives: Public Policy for Barred and Northern Spotted Owl," *Society & Animals,* 26: 217-238.
47. Marcus, A., & Oransky, I. (2018) "The data thugs," *Science,* 359: 730-732.

48. Martin v Waddell 1842. U.S. Supreme Court. 41 U.S. 367.
49. Miller, J., Stoner, K., Cejtin, M., Meyer, T., Middleton, A., & Schmitz, O. (2016) "Effectiveness of Contemporary Techniques for Reducing Livestock Depredations by Large Carnivores," *Wildlife Society Bulletin,* 40: 806-815.
50. Moreira-Arce, D., Ugarte, C. S., Zorondo-Rodríguez, F., & Simonetti, J. A. (2018) "Management Tools to Reduce Carnivore-Livestock Conflicts: Current Gap and Future Challenges," *Rangeland Ecology & Management*.
51. Mukherjee, S. (2010) *The Emperor of All Maladies: A Biography of Cancer*. Mew York: Scribner.
52. Mukherjee, S. (2016) *The Gene: an Intimate History*. New York: Simon & Schuster.
53. Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017) "A manifesto for reproducible science," *Nature Human Behaviour,* 1: 21.
54. Murtaugh, P. A. (2002) "On rejection rates of paired intervention analysis," *Ecology,* 83: 1752–1761.
55. OED. (2018) "Oxford English Dictionary," In. http://www.oed.com.ezproxy.library.wisc.edu/, 30 May 2017.
56. Ohrens, O., Bonacic, C., & Treves, A. (2019) "Non-lethal defense of livestock against predators: Flashing lights deter puma attacks in Chile.," *Frontiers in Ecology and the Environment,* 17: 32-38.
57. Open Science Collaboration. (2015) "Reproducibility Project: Psychology," *OSF*.
58. Pairs, B. (2018) "The cancer answer that wasn’t," In *Base Pairs*. Cold Spring Harbor Laboratory. https://www.cshl.edu/podcasts/base-pairs-episode-14-cancer-answer-wasnt/,
59. Platt, J. R. (1964) "Strong inference," *Science,* 146: 347–353.
60. Plous, S., & Herzog, H. (2001) "Reliability of Protocol Reviews for Animal Research," *Science,* 293: 608-609.
61. Plumer, L., Talvi, T. n., Männil, P., & Saarma, U. (2018) "Assessing the roles of wolves and dogs in livestock predation and suggestions for mitigating human-wildlife conflict and conservation of wolves," *Conservation Genetics,* https://doi.org/10.1007/s10592-017-1045-4.
62. Popescu, V. D., de Valpine, P., Tempel, D., & Peery, M. Z. (2012) "Estimating population impacts via dynamic occupancy analysis of Before-After Control-Impact studies," *Ecological Applications,* 22: 1389–1404.
63. Popper, K. (1959) *The Logic of Scientific Discovery*. Abingdon-on-Thames, U.K.: Routledge.
64. Prukop, J., & Regan, R. J. (2005) "In My Opinion: The value of the North American model of wildlife conservation—an International Association of Fish and Wildlife Agencies position," *Wildlife Society Bulletin,* 33: 374-377.
65. Robinson, M. (2005) *Predatory Bureaucracy*. Boulder, CO: University of Colorado Press.
66. Sanders, J., Jon Blundy, Anne Donaldson, Steve Brown, Rob Ivison, Miles Padgett, et al. (2017) "Transparency and openness in science," *Royal Society Open Science,* https://doi.org/10.1098/rsos.160979.

67. Santiago-Avila, F. J., Cornman, A. M., & Treves, A. (2018) "Killing wolves to prevent predation on livestock may protect one farm but harm neighbors," *PLoS One,* 10.1371/journal.pone.0189729.
68. Santiago-Avila, F. J., Lynn, W. S., & Treves, A. (2018) "Inappropriate consideration of animal interests in predator management: Towards a comprehensive moral code," In *Large Carnivore Conservation and Management: Human Dimensions and Governance,* edited by T. Hovardos. New York: Taylor & Francis.
69. Sax, J. L. (1970) "The Public Trust Doctrine in natural resource law: Effective judicial intervention," *Michigan Law Review,* 68: 471-566.
70. Shelton, M. (1973) "Some myths concerning the coyote as a livestock predator," *BioScience,* 23: 719-720.
71. Skloot, R. (2010) *The Immortal Life of Henrietta Lacks*. New York: Crown Publishers.
72. Stewart-Oaten, A. (2003) "On rejection rates of paired intervention analysis: comment," *Ecology 84, 2795–2799,* 84: 2795–2799.
73. Tompa, F. S. (1983) "Problem wolf management in British Columbia: Conflict and program evaluation," In *Wolves in Canada and Alaska: their status, biology and management,* edited by L. N. Carbyn. Edmonton, Canada: Canadian Wildlife Service, .
74. Treves, A. (2009) "Hunting to conserve large carnivores," *Journal of Applied Ecology,* 46: 1350-1356.
75. Treves, A., Artelle, K. A., Darimont, C. T., Lynn, W. S., Paquet, P. C., Santiago-Avila, F. J., et al. (2018) "Intergenerational equity can help to prevent climate change and extinction," *Nature Ecology & Evolution,* 2: 204-207.
76. Treves, A., & Bruskotter, J. T. (2014) "Tolerance for predatory wildlife," *Science,* 344: 476-477.
77. Treves, A., Chapron, G., López-Bao, J. V., Shoemaker, C., Goeckner, A., & Bruskotter, J. T. (2017) "Predators and the public trust," *Biological Reviews,* 92: 248-270.
78. Treves, A., Krofel, M., & McManus, J. (2016) "Predator control should not be a shot in the dark," *Frontiers in Ecology and the Environment,* 14: 380-388.
79. Treves, A., & Naughton-Treves, L. (2005) "Evaluating lethal control in the management of human-wildlife conflict," In *People and Wildlife, Conflict or Coexistence?,* edited by R. Woodroffe, S. Thirgood & A. Rabinowitz. Cambridge, UK: Cambridge University Press.
80. Treves, A., & Pizzagalli, D. (2002) "Vigilance and perception of social stimuli: Views from ethology and social neuroscience," In *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition.,* edited by M. Bekoff, C. Allen & G. Burghardt: MIT Press, Cambridge, MA.
81. Treves, A., Santiago-Ávila, F., & Lynn, W. S. (2018) "Just Preservation," *Biological Conservation,* 229: 134-141.
82. USDA-WS, & others. (2016) "Final Environmental Assessment: Predator Damage Management in Idaho," In: U.S. Department of Agriculture.
83. van Eeden, L. M., Ann Eklund, Jennifer R. B. Miller, José Vicente López-Bao, Mikael R. Cejtin, Guillaume Chapron, et al. (2018) "Carnivore conservation needs evidence-based livestock protection," *PLOS Biology,* https://doi.org/10.1371/journal.pbio.2005577: https://doi.org/10.1371/journal.pbio.2005577.

84. van Eeden, L. M., Crowther, M. S., Dickman, C. R., Macdonald, D. W., Ripple, W. J., Ritchie, E. G., et al. (2018) "Managing conflict between large carnivores and livestock," *Conservation Biology,* doi: 10.1111/cobi.12959.
85. Vucetich, J. A., Burnham, D., Macdonald, E. A., Bruskotter, J. T., Marchini, S., Zimmermann, A., et al. (2018) "Just conservation: What is it and should we pursue it?," *Biological Conservation,* 221: 23–33.
86. Vucetich, J. A., J.T., B., Nelson, M. P., Peterson, R. O., & Bump, J. k. (2017) "Evaluating the principles of wildlife conservation: a case study of wolf (Canis lupus) hunting in Michigan, United States," *Journal of Mammalogy,* 98: 53-64.
87. Wadman, M. (2018) "Hundreds of U.S. scientists urge more transparency in animal research," *Science*: oi:10.1126/science.aau5395.
88. Wagner, K. K., & Conover, M. R. (1999) "Effect of preventive coyote hunting on sheep losses to coyote predation," *Journal of Wildlife Management,* 63: 600-612.
89. Walsh, E., Rooney, M., Appleby, L., & Wilkinson, G. (2000) "Open peer review: a randomised trial," *British Journal f Psychiatry,* 6: 47-51.
90. Washington, H., Chapron, G., Kopnina, H., Curry, P., Gray, J., & Piccolo, J. (2018) "Foregrounding ecojustice in conservation," *Biological Conservation,* 228: 367-374.
91. Western Watersheds Project et al. v USDA Wildlife Services. 2018. U.S. District Court for the District of Idaho. 1:17-cv-00206-BLW Document 22-3.