Appendix to ‘Reducing meat and animal product consumption: what works?’

1. A few comments on our inclusion criteria

a. Our inclusion criteria are motivated by a previous project of Seth’s, which found that assessing the effects of intergroup contact on prejudice, a meta-analysis that looked at absolutely everything came to much more optimistic conclusions than one that looked just at the RCTs with at least one day separating treatment and outcome measurement. I (Seth) had an intuition that we’d find something similar in the MAP reduction literature, and I proposed a project investigating this to Benny when we met at EAGxNYC in August 2023. (Our first interaction was on the EA forum.)

b. The standards pertaining to sample sizes are adapted from guidance in Paluck et al. (2021) (on which Seth worked as an RA), which found that anti-prejudice studies with fewer than 25 subjects in treatment and control tended to produce systematically larger effects, and argued that for cluster-assigned studies, “estimating standard errors requires at least 10 clusters.”

c. Regarding our focus on MAP consumption as the key outcome: attitudes and behavioral intentions are interesting as well, and a future project might investigate the link between attitudes/intentions and behavioral change. But for now, we take our cues from Mathur et al. (2021b), who find a substantial “discrepancy between intentions and reported consumption” and conclude that “reported intentions may be a poor proxy for reported actual consumption” (this is a well-understood problem in dietary, health, and psychology research).

d. Regarding the delay criterion: for some in-restaurant studies, e.g. Piester et al. (2020), any individual participant will have their outcomes measured the day that treatment was administered; in such cases, we looked for studies where either the intervention or outcome measurement period (or both) spanned multiple days.

e. A more conceptual note on the benefits of a focused meta-analysis: meta-analyses often have a “garbage in, garbage out” problem. When you average the results from three unreliable studies and one good one, you often end up with an average effect size that seems stronger and more precisely estimated than any individual study under review, but in a way that papers over structural weaknesses in the papers. The metaphor that comes to mind is packaging up a bunch of subprime mortgages together into a tranche and calling it AA. Others argue that this method effectively compares apples and oranges — and sometimes “apples, lice, and killer whales.”

2. Self-reported outcomes may induce social desirability bias

Mathur et al. (2021b) argue that social desirability bias is "widespread" in the MAP reduction literature. While our meta-analysis did not find systematic differences in effect sizes between studies that measured MAP consumption either obliquely or through self-report,, we still think that Mathur et al. are basically right because for some of these studies, it's almost impossible to imagine that participants don't know the answer that researchers are hoping to hear. Consider Fehrenbach (2015), a 2015 dissertation that tested the "effectiveness of [two] video messages designed to encourage Americans to reduce their meat consumption." The study had two treatment arms and a control. Both treatment videos sought to induce a feeling of "high threat" by informing viewers of the "the negative health effects of high meat consumption;" one video also sought to induce feelings of "high efficacy" by suggesting "easy ways to reduce their meat consumption," while the "low efficacy" group's video "only included a very minor efficacy component in the conclusion." The videos were 7 and 4 minutes long, respectively. Before the study, on the day of the study, and again a week later, participants were asked about their attitudes and intentions towards eating meat, as well as how much meat they'd eaten in the past 7 days.

Overall, the high threat/high efficacy group reported that they ate an average of 3.16 fewer meals involving meat in the week following the intervention than the one before it, compared to 2.11 for the high threat/low efficacy group and 1.92 for the control group. As a benchmark, the population ate meat at an average of 13.64 meals per week before the intervention (SD = 4.21).

This study strikes us as having a high risk of social desirability bias for three reasons.

First, the study is designed to make people feel a sense of "high threat" from eating meat, and then asks them a week later about how much meat they ate. There are grounds for doubting how much respondents would accurately recount their eating habits. This problem is typical of this literature.

Second, the study asks participants to recall a week's worth of meals; previous research has found that daily food diaries lead to more accurate reports. As Mathur et al. (2021a) put it:

> Many existing studies measure meat consumption in terms of, for example, Likert-type items that categorize the number of weekly meals containing meat (e.g., "none", "1–5 meals", etc.) or in terms of reductions from one's previous consumption. When possible, using finer-grained absolute measures, such as the number of servings of poultry, beef, pork, lamb, fish, etc., would enable effect sizes to be translated into direct measures of societal impact.

Third, the decline in meat-eating among the control group suggests that the intended direction of the experiment might have been crystal clear to everyone, whether they watched the video or not.

In sum, using broad stroke, self-reported outcomes in a context where meat is being presented as bad for you seems like a high-risk environment for experimenter demand effects.

3. Why publication bias is a problem, and a method for detecting it

One of the most useful things a meta-analysis can add to our understanding of a literature is an assessment publication bias, sometimes called the file drawer problem. The concern is that many academic disciplines have a bias towards positive, significant findings, so when researchers test a hypothesis and find null results, they're more likely to either shelve the study or be unable to publish it. In this particular dataset, the fact that only 20 of 40 interventions report significant effects is already reasonably strong evidence that this particular literature (or this particular outcome *within* the experimental literature) doesn't have that problem. But another way to test for publication bias is to see if more precise studies are systematically more likely to report smaller effects.

There are a few ways to think about why you might observe such a pattern in the presence of publication bias. One is that a more precisely estimated study (one with a smaller standard error) is a lot more likely to be closer to the true effect size by dint of its precision; so if big studies show small results and small studies show big results, that's straw-in-the-wind evidence that the small studies are drawn from a truncated distribution where small, insignificant results are more likely to be shelved. Another way to think about a potential relationship between effect size and precision is that it's a lot easier/more tempting to shelve a small study than a big one. If you run a small study and find nothing, you might write it off and move on. But big studies take a lot of work — imagine how many hours went into a study reaching 9 million people in 516 treated areas — and researchers might be more inclined to try to get *something* out of a big study, i.e. see it through to publication come hell or high water.

4. Some methodological issues and trends we encountered while coding studies

4.1 There should be a control group and it shouldn't receive pro-vegan messaging

Four interesting studies did not meet our inclusion criteria because of issues with their control groups (one, two, three, four). Either they didn't have one or the control group received some form of vegan messaging, e.g. people were assigned to treatment or control in a tent plastered with animal rights imagery. But the control group should be a control group.

That also means not to try to *over*-control their behavior. For instance, one study's control group "were asked not to change their diet" in any way. But *that message is a kind of treatment*. We understand the instinct to try to limit noise, but that can't come at the expense of internal validity. A control group that gets told to do or not do something by experimenters has received an

experimental stimulus. This violates the equivalence in expectation assumption that RCTs depend on for unbiased inference.

4.2 The garden of forking paths

Cooney (2014), "What elements make a vegetarian leaflet more effective?" is a Humane League study that looked at how to make a "pro-vegetarian booklet more effective at inspiring young people to reduce their consumption of animal products." The study recruited young people at colleges and universities in the Northeastern U.S. and at various stops on the Warp Music tour (a wonderful detail!) and randomly assigned them to read one of eight booklets or a control group. The booklets presented (some combination of) information about animal cruelty, the health benefits of going vegetarian, information focused on all farm animals or just chickens, and content on either **why** participants should go vegetarian or a focus on **how** to do so. 3,233 people filled out an initial survey on their eating habits and viewed one of the booklets (or were assigned to control); 569 people filled out an endline survey 3 months later to assess long-term impacts.

The study's main findings are that information on cruelty was slightly more effective at reducing MAP consumption than health information was; that leaflets focused on *all* farm animals outperformed those focused just on chickens; and information about how to go vegetarian was more effective than arguments for going vegetarian. Overall, Cooney concludes, the study produces "strong evidence that pro-vegetarian booklets aimed at young people should focus on all main farm animals, not just chickens," and "weak evidence that booklets should focus more on the cruelty done to animals than on the health benefits of going vegetarian;" once these two elements are established, he argues, it's "unclear whether how-or why-focused booklets will be more effective at persuading people to change their diet."

However, each of these is an intra-treatment comparison, and ignores the control group. For a meta-analysis, the main quantity of interest is generally how a given treatment, or cluster of treatments, compares to baseline of no intervention. For these participants, three months later, "those in the control group (those who never received a booklet) reported more of a reduction in animal product consumption than those who received any of the other booklets." Coney does not believe that the booklets actually reduced the likelihood of changing one's diet, and calls this an anomaly. He also partly attributes it to tiny sample sizes.

This is a textbook example of the garden of forking paths. Choosing an analytic strategy after you've seen the data allows researchers to make *prima facie* reasonable choices that, at the aggregate level, have a way of suppressing and under-reporting backlash or unwelcome results.

This study would have benefited from preregistration and a clear divide in its analyses between hypotheses established *a priori* and those developed after data collection. However, and to the team's credit, the study's code and data were posted to the Open Science Framework, and interested readers can reconstruct whatever analyses they think most relevant.

4.3 Social psychology studies were sometimes hard to read and parse for statistical information

The psych studies in this literature are often drawing from theories that are unfamiliar to us. For instance, Fehrenbach (2015) is a test of the "Extended Parallel Process Model," which describes "how rational considerations (efficacy beliefs) and emotional reactions (fear of a health threat) combine to determine behavioral decisions." This study had treatment subjects watch a video that sought to induce a feeling of "high threat" by describing the harms and dangers of meat, while also giving some subjects a sense of "high efficacy," i.e. the sense that they are capable of altering their behavior. We don't know anything about this strand of social psychological theory. However, for our purposes what ultimately matters is whether these interventions reduce MAP consumption or not.

We also sometimes found the analyses presented in these studies hard to parse. In lieu of a difference of means, these studies often reported statistics that combined the 'effects' of the treatment itself and non-randomly assigned moderators (e.g. gender) or mediators (e.g. a score on a personality test trying to diagnose people's reasons for resisting vegetarianism). These approaches have notable conceptual drawbacks, but for our purposes, the main drawback was that they added a lot of additional uncertainty to the process of figuring out the *average* treatment effect: the difference in means between the treatment and the control groups at posttest. We ask that researchers in this field please heed Daniel Lakens's observation that any given study "is just a data-point in a future meta-analysis," and make future researchers' lives easier by clearly documenting five essential pieces of information: the control group's sample size and mean posttest score, the treatment group's sample size and mean posttest score, and a measure of sample variance, rather than estimator variance, i.e., the standard deviation rather than (or in addition to) a standard error. (These two quantities are sometimes confused in meta-analyses.) Please and thank you!

4.4 If you cluster, please tell us how many units per cluster

Hennessey (2016) tells us that the average cluster has 1.6 participants, whereas Piazza et al. (2022) say that treatment was assigned to people in groups of 2-8. We therefore estimated the number of clusters by dividing the Ns by 5 (the numeric mean of 2-8). But this isn't very precise. It's not a huge risk of bias, but in general, we ask researchers to please make future meta-analysis as easy as possible by reporting information like this as clearly as they can.

4.5 Measure at whatever level of fidelity makes sense

Gravert and Kurz (2019) look at meat, fish and veggies separately. Likewise, Jalil et al. (2023) measure red meat and other kinds of meat separately. This reporting clarity enables all kinds of subsequent re-analyses and theory-building. If at all possible, we ask researchers to report granular data about categories of MAP consumption.

5. Summaries of prior reviews

Here are our notes on three prior reviews: Bianchi et al.(2018a) on the "conscious determinants" of MAP consumption; 2) Bianchi et al. (2018b) on changes to the microenvironment; and 3) Mathur et al. (2021a) on appeals to animal welfare.

5.1 Bianchi, Dorsel, Garnett, Aveyard & Jebb (2018): "Interventions targeting conscious determinants of human behaviour to reduce the demand for meat: a systematic review with qualitative comparative analysis"

This paper reviewed interventions targeting "conscious determinants" of demand for meat, and found "24 papers reporting on 29 studies" with 37 interventions.

These Interventions were broadly grouped into information about

a. meat consumption and health (N = 11)
b. the environment (N = 8)
c. animal welfare (N = 2)[1]
d. socio-economic issues (N = 2)
e. or a combination thereof (N = 14).

Of the 29 studies, 15 reported "actual meat consumption," 6 looked at "meat purchase or selection" (hypothetically), and 15 looked at "intended consumption." (Some studies look at multiple dependent variables.)

Overall, the authors find that "self-monitoring and individual lifestyle counselling interventions showed promise in reducing actual consumption of meat." However, while education about the health, environmental, and animal welfare consequences of eating meat showed promise in changing *intended* behaviors, they did not appear to change actual MAP consumption.

5.2 Bianchi, Garnett, Dorsel, Aveyard and Jebb (2018): "Restructuring physical micro-environments to reduce the demand for meat: a systematic review and qualitative comparative analysis"

This paper reviewed the choice architecture literature and found "14 papers reporting on 18 studies with 22 intervention conditions" that aim to reduce "demand for meat, defined as the actual or intended consumption, purchase, or selection of meat in real or virtual environments." Of the 18 included, the authors consider 3 to have "strong" methodological quality, 11 to have medium, and 3 to be of low quality. "Six studies reported data on meat consumption," while the remaining 12 report on either meat purchases/selection or meat purchases/selection in a virtual setting.

[1] Three years later, Mathur et al. (2021a) found 34 papers in this category.

Those 18 studies fell into 7 bins, and found the following results:

- The three studies that "reduced the portion size of meat servings in restaurants or laboratory settings…significantly reduced meat consumption." These studies literally served some people less meat than others and then observed how much meat everybody ate, e.g. Rolls, Roe, and Meengs (2010). It would have been shocking to us if they hadn't found an immediate effect; but none measured lasting changes.
- The three observational studies that "provided meat alternatives to freeliving individuals (ie, those not being observed in a laboratory setting)…were associated with significant reductions in meat purchases or consumption."
    - The two of these we could find online (Flynn, Reinert and Schiff 2013, Holloway, Salter, and McCullough 2012) both implement very nice, behaviorally-based interventions that would benefit from randomized replication.
- The four studies that "altered the visual aspects or the hedonic appeal of meat or meat alternatives…significantly reduced the demand for meat in virtual food choices." All of these took place in virtual settings, meaning they didn't measure actual eating habits.
- The four studies that "repositioned meat products to reduce their prominence at point of purchase" found mixed results. Two of these studies "reduced or were associated with reductions in meat demand," while two others did not see statistically significant results on meat demand.
- The five studies that "manipulated menus and meal booking systems by changing the verbal description or label of meat or meat alternatives" had differential results by study design. A non-randomized study that altered "university meal booking systems to refer to meat options as 'meat' rather than 'standard' or 'normal' was associated with reduced meat purchases;" however, four randomized studies that changed verbal descriptions of items on a menu did not find an effect.
- The one study that "used a pricing intervention" — basically it created a constant price per ounce for differently sized orders of chicken nuggets, as opposed to pricing the largest order at a lower price by weight — found no effects on meat consumption "in a simulated food choice task."
    - The study surveyed people *at a fast food restaurant* but did not measure any actual food choices. This is baffling.
- Last, the two studies that "changed multiple elements of a university canteen or of small businesses" found mixed results. A study of a marketing campaign at a university providing "examples of meatfree dishes at the canteen entrance, indicators of healthy meat free options, and educational flyers" reduced meat consumption. However this study was badly underpowered, with just two units in treatment and two in control. Meanwhile, an "18 month multicomponent intervention targeting red meat consumption and other health behaviours" at 18 worksites did not meaningfully reduce meat consumption; further, its specific changes to the microenvironment "were not reported in detail, precluding more detailed analyses of this intervention."

One noteworthy field experiment in this literature is Sorensen et al. (2005), which tested an intervention "designed to improve health behaviors among working-class, multiethnic populations employed in small manufacturing businesses," and found no meaningful change in

red meat consumption. Of the remaining ten studies we checked, five measured hypothetical meat choices rather than actual consumption, two featured no delay, two did not have enough clusters to qualify, and one was too barebone to draw any quantitative results from (though it too reported null results).

5.3 Mathur, Peacock, Reichling, Nadler, Bain, Gardner, and Robinson (2021): Interventions to reduce meat consumption by appealing to animal welfare: Meta-analysis and evidence-based recommendations

This paper meta-analyzes 34 papers, comprising a total of 100 recorded outcomes, that attempt to reduce MAP consumption through appeals to animal welfare. These appeals are typically undergirded by one or more of nine psychological theories: 1) changing social norms; 2) the "identifiable victim effect" where people typically have a stronger reaction to something bad happening to a named individual rather than a class or group; 3) implementation suggestions, e.g. concrete ideas for meat substitutes at breakfast; 4) meat-animal reminders that link meat to actual creatures; 5) "moral shock" tactics that highlight how horrible factory farming is; 6) mind attribution, e.g. prompting people to imagine the lived experience of a cow; 7) the "moral equivalence of farm animals and companion animals;" 8) linking physical and moral disgust; and 9) the opportunity to participate in, and identify with, a social movement.

For whatever reason, only 24% of these articles were published in peer-reviewed journals; the remainder came from "dissertations, theses, conference proceedings, or reports by nonprofits." Also notable is that ⅔ of the interventions took less than 5 minutes to implement. Of the 100 point estimates,75 come from randomized studies, 96 were self-assessments of MAP consumption, and 53 were measured immediately following the intervention.

Pooling all studies together, the authors find that the interventions collectively led to a 22% increase in "subject's probability of intending, self-reporting, or behaviorally demonstrating low versus high meat consumption, compared to the control condition." In other words, a meaningful average effect.

Overall, this results suggests that appeals to animal welfare generally work at changing behavior. However, a close look at the reported data reveals a few grounds for doubt.

1. Only 25 of these point estimates were both in the less-MAP direction and statistically significant outcomes; the remaining 75 either were nonsignificant or in the wrong direction.
2. Looking at the ten largest effect sizes, nine are intended behavior outcomes, and the tenth (Norris and Roberts 2016) finds big results, but is not an RCT because it has no control group.[2]
    a. Treatment *was* randomly assigned, just to different treatment groups. However, because the results from each treatment group were broadly similar, the authors pool all results into one random effects outcome and treat it as a pre-post

[2]

outcome. We further note that the study was conducted by an advocacy group rather than independent researchers, was never published in a journal, and doesn't provide any information on the participants besides the fact that they were mTurk participants. Incorporating these kinds of data points into our assessment of a study's evidentiary value is tricky. Our main concern is the lack of a control group.

3. A delay of 7 days is associated with a curtailing of the pooled effect size to zero. The risk ratio associated with studies with a delay of at least 7 days is 0.81, with the original effect of 1.22 set as baseline; 0.81 * 1.22 = 0.9882, which means that the treatment and control groups have effectively equivalent outcomes (or slightly lower in the treatment group).
4. The 21 point estimates from pre-registered studies with open data have a pooled risk ratio of 1.09, an estimate which falls just short of the conventional standard for statistical significance ($p = 0.06$). We're not sticklers for this or any particular significance threshold; but when the best studies in a dataset find systematically smaller results, it suggests that a literature's effect sizes might partly be a function of researcher degrees of freedom.
5. Finally, these are all self-reported results; as the authors put it, this might bias results if it "induce[s] misreporting that is *differential* between the intervention and control group, which could potentially inflate estimates away from the null."

Overall this is a great meta-analysis and we recommend reading it if you enjoy the subject.