Taxonomic disagreement about ranks in gray-area taxa: A vignette study

Peer-reviewed author version

Conix, Stijn; CUYPERS, Vincent; Zachos, Frank E. & De Block, Andreas (2023) Taxonomic disagreement about ranks in gray-area taxa: A vignette study. In: BIOSCIENCE, 73 (10) , p. 728 -737.

DOI: 10.1093/biosci/biad081
Handle: http://hdl.handle.net/1942/41652

Taxonomic disagreement about ranks in grey-area taxa: a vignette study

Stijn Conix[1,3], Vincent Cuypers[2,3], Frank E. Zachos [4,5,6] & Andreas De Block[7]

[1] Corresponding author Institut Supérieur de Philosophie, Université Catholique de Louvain, Collège Désiré Mercier, Pl. Cardinal Mercier 14, 1348 Louvain-La-Neuve, Belgium, 0000-0002-6836-6047 stijn.conix@uclouvain.be

[2] Research Group Zoology: Biodiversity and Toxicology, Centre for Environmental Sciences, Hasselt University, Agoralaan Gebouw D, B-3590 Diepenbeek, Belgium, 0000-0002-9556-5359, vincent.cuypers@uhasselt.be.

[3] Centre For Logic and Philosophy of Science, KU Leuven, Kardinaal Mercierplein 2, 3000 Leuven, Belgium.

[4] Natural History Museum Vienna, Burgring 7, 1010 Vienna, Austria, 0000-0002-0133-6265, frank.zachos@NHM-WIEN.AC.AT.

[5]Department of Genetics, University of the Free State, Bloemfontein, Park West, Bloemfontein, 9301, South Africa

[6] Department of Evolutionary Biology, University of Vienna, Djerassiplatz 1
A-1030, Vienna, Austria

[7] Centre For Logic and Philosophy of Science, KU Leuven, Kardinaal Mercierplein 2, 3000 Leuven, Belgium, 0000-0002-7927-8210, andreas.deblock @kuleuven.be.

Abstract

When producing species classifications, taxonomists are often confronted with grey-area cases. It has been claimed that in such cases, the ranking decision is in part subjective and may differ between taxonomists due to differences in species concepts or even conservation values. Here, we

use a vignette study to test this claim, and explore the drivers of taxonomic decision-making. For three fictional taxonomic scenarios, we asked the opinion of taxonomists on one of multiple versions of an abstract containing a ranking decision. The cases were designed to represent grey-area cases, and differences between versions related to potential drivers of decisions. Our results suggest that taxonomists tend to disagree moderately about species-ranking decisions in grey-area cases even when presented with the same data. We did not find evidence that species concepts or conservation values are drivers of taxonomic disagreement. Instead, the use of different kinds of data seemed to be more important.

Keywords: grey-area taxa; species delimitation; methods in taxonomy; vignette study; taxonomic disagreement

Introduction

For many taxa across the Tree of Life, specialists in taxonomy disagree about how to classify them. Such disagreements often revolve around the rank of groups, for example whether they should be recognized as species or as subspecies. This is often explained by the fact that taxonomic classifications enforce a binary system – a group of organisms either is recognized as a species or not – onto differences that are usually gradual and continuous, rather than discrete (Zachos et al. 2020, Thiele et al. 2021). Many of the criteria in use for delimiting species, such as morphological or molecular distinctness, interfertility or ecological niche differentiation, indeed apply to groups of organisms in various degrees. While many groups are clearly distinct, probably warranting species status, other groups find themselves in a 'grey area' between what are typically accepted as good separate species and what are not. The appropriate ranking decision in such grey-area cases is not clear-cut, and taxonomists may disagree even if they use the same data and criteria.

Because ranking decisions in grey-area cases often diverge among taxonomists, some have called ranking decisions in taxonomy at least partly 'subjective' (Mishler and Wilkins 2018, Zachos et al. 2020, Zachos 2022). If taxonomy is indeed subjective in that way, that could pose problems for the discipline. Not only could it fuel unnecessary debates in a discipline that already lacks funding and researchers, but it would also affect all scientific and non-scientific domains that rely on the species-level classifications that taxonomists generate (see e.g. Faurby et al. 2016, Willis 2017). These domains typically assume that all units at the species level are similar and, importantly, directly comparable, but, if ranking decisions are sometimes executive decisions rather than completely evidence-based, this assumption may be unfounded. Disagreements often lead to the circulation of competing classifications (McClure et al. 2020, Neate-Clegg et al. 2021), which results in different groups of users using different classifications – making synergizing efforts often difficult – and forces users to invest in taxonomic decision-making themselves.

In order to reduce the disorder that disagreements and uncertainty create in taxonomy, it is important to know what drives taxonomists' decisions and disagreements in grey-area cases. Three main factors are regularly cited as the determinants of disagreement at this level. First, it is commonly claimed that taxonomists' preferred species concept(s) could explain why their conclusions differ from those of colleagues. In particular, it is often assumed that taxonomists preferring the Biological Species Concept (BSC) are less likely to split taxa into smaller groups than taxonomists who prefer the diagnosability version of the Phylogenetic Species Concept (dPSC) (Agapow et al. 2004, Isaac et al. 2004).

Second, some have argued that taxonomic disagreement about grey-area cases is driven by differences in the way species concepts are methodologically operationalized (Camargo and Sites 2013, Satler et al. 2013, Conix 2018). In the case of the BSC and the dPSC, this translates into debates about the importance of gene flow and (cryptic) molecular differentiation, and the various ways in which an abstract notion such as gene flow can be tested in practice.

Finally, and more controversially, some have claimed that taxonomists are sometimes influenced by non-taxonomic considerations such as the implications of ranking decisions for conservation (Karl and Bowen 1999, Isaac et al. 2004). Here, the claim is usually that taxonomists are more likely to recognize threatened groups as distinct taxa (species or subspecies) hoping that this would improve chances of legal protection or conservation action for those groups. Other value-laden factors that potentially play a role include economic, political and sociological factors. It is commonly believed, for example, that there were strong lumping traditions in both bird and mammal taxonomy in the past (Cotterill et al. 2014, Sangster 2014). Similarly, more experienced taxonomists and taxonomists working in low-income countries may rely more on morphological evidence (as an operationalization) than young taxonomists and taxonomists working in high-income countries. Taxonomists working in countries with relatively low diversity, on the other hand, may have a stronger tendency to split (Harris and Froufe 2005).

To our knowledge, these explanations have never been experimentally tested. While there have been at least two surveys on which species concepts biologists use and how they use them (Pušić et al. 2017, Stankowski and Ravinet 2021), these were simple self-report surveys among biologists of many subdisciplines. This study, instead, is experimental and only included responses from practicing taxonomists. Our aims are 1) to test whether taxonomists indeed sometimes make different ranking decisions given the same data (subjectivity), and 2) to investigate what kinds of taxonomic and non-taxonomic information, particularly species concepts, evidence types (operationalization) and conservation values, are most likely to influence ranking decisions. To accomplish this, we carried out an online vignette study in which respondents were asked to evaluate three fictional taxonomic cases. For each case, any single respondent was presented with one of multiple slightly differing versions of the same abstract, and had to state whether they agreed with the decisions made in that abstract. This allowed us to quantify variation in the responses of taxonomists in general, and variation between groups of taxonomists that had received a different version of the abstracts.

Methods

This study was approved by the Social and Societal Ethics Committee of KU Leuven, Belgium (file G-2022-4955-R2(MIN)). Apart from country of residence, no personal data were collected, and the data were published with country of residence aggregated into continents and low/high income country (as classified by the World Bank (2022)), to guarantee anonymity of the respondents. Data collection only started after the full research design was preregistered on the Open Science Framework. The full questionnaire, analysis plans, raw data, analysis code and supplementary materials can be found on the Open Science Page (Conix et al. 2022) of the research project.

Design

We designed an online survey consisting of questions about respondent characteristics and three fictional taxonomic abstracts ('vignettes'). The respondent characteristics included whether the respondent is a taxonomist, whether they do this professionally, their experience, country of residence, taxon of specialization, whether they read taxonomic literature outside of their area of expertise, and their preferred species concept. To avoid influencing responses to the vignettes, respondents were asked about their preferred species concept only after evaluating the vignettes.

Each participant was given three vignettes in random order. These included one abstract describing a new fictional plant species, one abstract describing a new fictional frog species, and one abstract describing a new fictional flatworm species. All three fictional groups were designed to be grey-area cases. We chose to use fictional taxa to avoid that taxonomists' pre-existing opinions on real taxa would influence their decision. We chose a plant, frog and flatworm in order to have at least one taxon that respondents would be likely to know little about (the flatworm), one taxon that respondents are likely to be somewhat familiar with (the frog), and one non-animal case (the plant).

For each vignette, there were several versions that we designed to differ as little as possible apart from the condition under investigation. The plant case was designed to investigate the role of conservation values. We included a version of the vignette stating that the taxon is threatened, a version stating that the taxon is not threatened, and a neutral version with no information about the conservation status. The frog case was designed to test the role of operationalization, and the different versions differed in the kinds of evidence types they included. The neutral version only included limited morphological data. The other versions added more morphological data, mtDNA data and ecological data, respectively. The flatworm case centered around gene flow, with a version mentioning gene flow, a version mentioning the absence of gene flow, and a neutral version not mentioning gene flow at all. Because gene flow is tightly related to reproductive isolation, this vignette served as a test of the influence of species concepts on ranking decisions. See table 1 for an overview of all cases and conditions.

Each respondent was randomly assigned one version of each of the three cases. For each case, respondents were asked whether they agreed with the ranking decision (i.e. with the proposed new species) in the abstract. This is the main outcome variable of the study. For the frog case, they were also asked which kind of evidence they thought was lacking in case they did not agree with the ranking decision in the abstract. Each respondent was also asked whether they would accept the abstract for a conference presentation. This question was included to check whether respondents perceived the abstracts as scientifically legitimate. The full survey, with vignettes, is available in the supplementary materials S1.

Sampling

The survey was distributed through several taxonomic mailing lists with one reminder two weeks after sending out the survey. In addition, the survey was disseminated through the networks of the authors and sent to various professional organizations and natural history museums asking to disseminate the survey (for a full list of organizations and institutions contacted, see supplementary materials S2). Because the number of responses from outside of Europe was low after the full preregistered sampling period, we kept the survey open for one month longer than initially planned, and sent the survey out through our networks in South America, Africa and Southeast Asia. Because we expect the sample to be more representative of the population after the additional sampling effort, all exploratory analysis reported below was done using the extended dataset. Because the additional sampling was not preregistered, the registered hypothesis tests were done using the original, smaller dataset.

In order to use only high-quality data for the analysis, we only retained responses that took at least 150 seconds to finish the survey (i.e. the minimum time needed to read and process all questions) and responses that replied to at least one of the three main outcome questions. We also only retained responses from respondents who indicated that they are taxonomists.

It is likely that some participants received the invitation for the survey more than once because of overlaps between different channels of dissemination. Because of the snowball method of sampling, it is not possible to estimate the response rate and difficult to estimate how representative the sample is of the wider population.

Statistical analysis

We registered two hypotheses for this study on OSF: (1) If taxonomists frequently make different ranking decisions even when given the same data, then there will be strong disagreement about the ranking decision for the three abstracts across conditions. (2) If taxonomists are more likely to rank a group as a species if it is threatened, then the proportion of 'agree' will be substantially higher for participants assigned to the 'threatened' condition than for participants assigned to the 'not threatened' or 'neutral' condition.

The aim of the first hypothesis was to establish that taxonomists indeed make divergent ranking decisions, even when given the same data. While this may be obvious to most working taxonomists (Isaac et al. 2004, Tattersall 2007, see e.g. Heller et al. 2013), it has, to our knowledge, never been experimentally quantified. We tested this hypothesis using a simple Bayesian model to estimate the proportions of 'agree' for all conditions for each of the three cases. We registered in advance that we would consider there to be strong disagreement about an abstract in case the entire 0.80 highest density interval (hdi) of the estimated proportion of the minority opinion for that abstract is above 0.25. This would mean that it is highly likely that at least 25% of taxonomists would have a different opinion about the ranking decision in that abstract than the majority opinion.

The aim of the second hypothesis was to test whether, as some have claimed (Isaac et al. 2004, Conix 2019), non-taxonomic considerations such as conservation values influence ranking decisions. To test this hypothesis, we estimated the proportion of 'agree' for the 'threatened' and 'abundant' condition in the plant case using a Bayesian model, and subtracted the posterior distributions of these

proportions. We stipulated in advance that we would accept the hypothesis if 0 would fall outside of the 0.80 hdi of the resulting distribution (meaning that the estimated difference between the two conditions is highly unlikely to be 0).

In addition to these two registered hypothesis tests, we designed a causal model incorporating the main factors cited in the literature as potential causes of disagreement (figure 1). This causal model is based on our experience of the field and the literature on the species problem, and could serve as a basis for future research into the causes of ranking disagreement. The model shows that we assume that ranking decisions are influenced by the species concept of taxonomists, the particulars of the case (i.e. the treatments), but also by which taxon they specialize in and whether they are active in a low-income country. The latter two factors capture, among other things, the research community and research culture taxonomists are active in, and influence ranking decisions both directly and through an influence on their species concept.

We used this causal model to select predictors for regressions testing the causal role of conservation values in the plant case (model 1), operationalization in the frog case (model 2) and species concepts in the flatworm case (model 3). More precisely, we applied the so-called 'backdoor criterion' to the causal model to select the variables to condition on and avoid that the estimates would be influenced by non-causal paths in the model (Cinelli et al. 2022). In addition, we tested the conditional independencies of this model where possible to ascertain there were no strong associations between variables where the model did not predict this to be the case (Cinelli et al. 2022). Note that these models were exploratory, and designed after collecting and seeing the data. Hence, even though they were designed using a causal model, they should be interpreted with caution and mostly as the basis for designing further, ideally preregistered hypothesis-driven research.

All three models were Bayesian logistic regressions with 'agree' as the outcome, 'treatment' and 'species concept' as predictors of interest, and taxon of specialization and income status of the home country (high or low) as control variables. They all included a general intercept and offsets for

the included groups (income, species concepts, taxon of specialization) and treatments (conditions of the vignettes). In model 1, we also included a varying effect of treatment by income status, as we expected that the influence of conservation values might differ by low/high income. Similarly, we included varying effects for both taxon of specialization and income status by treatment in model 2, as we assumed that the influence of different operationalizations could differ by taxonomists' specialization and whether they worked in a low-or high-income country. Finally, we included a varying effect for species concept by treatment in model 3, as we expected that the influence of the gene-flow condition might differ depending on the species concept participants subscribed to.

Only participants who responded to all questions included in the analysis (agree, income status, taxon of specialization, species concept) were included in the analysis (N = 423). We used weakly informative priors in all these regressions, and all the analyses were accomplished using Markov chain Monte Carlo methods (van Ravenzwaaij et al. 2018). For a full specification of the models as well as the Pymc code used to run them, see supplementary materials S3. For an overview of the three models, see table 2. We used Pandas (McKinney 2010), Scipy (Virtanen et al. 2020) and Numpy (Harris et al. 2020), Seaborn (Waskom et al. 2022) and Matplotlib (Hunter 2007) in a jupyter notebook for all descriptive analyses. We used the Pymc (Salvatier et al. 2016), Bambi (Capretto et al. 2022) and Arviz (Kumar et al. 2019) libraries in python (see source code for all package versions) for all hypothesis tests and exploratory regressions.

Results and discussion

After both sampling periods, the survey was filled in by 706 participants. After removing responses that took less than 150 seconds, responses without answers to the main outcome questions ('Do you agree?'), and responses of participants that indicated they were not taxonomists (97 in total), 447 responses were left. This is substantially more than in previous surveys on species concepts (Pušić et al. 2017, Stankowski and Ravinet 2021), in particular if only taxonomists are considered.

For the two hypothesis tests, we removed the responses received after the preregistered sampling period (N=51) as well as responses with missing data for one the variables included, keeping 396 responses for hypothesis 1 and 389 for hypothesis 2.

The main respondent characteristics for this dataset of 447 respondents are summarized in tables 3 and 4, and visualized in supplemental figure S4.1). Respondents were relatively equally divided between various species concepts. The diagnosability version of the Phylogenetic Species Concept (dPSC, 28.3%) was most popular, closely followed by the Biological Species Concept (BSC, 24.5%) and the Evolutionary Species Concept (ESC, 24.1%). It is notable that 43.6% of our sample was European, 32.8% of our sample has worked for at least 30 years since their PhD, and the distribution of specializations was in most cases (but not always) clearly different from the species richness of the taxa they specialize in. This is in line with another relatively large survey among taxonomists (Salvador et al. 2022) and our expectation that western taxonomists, taxonomists in senior positions and taxonomists working on vertebrates (and insects) make up a relatively large share of all taxonomists. The distribution of specializations is also broadly in line with the proportions of mentions of taxa from these groups in a large full-text corpus of taxonomic research papers (see Pence and Conix Forthcoming). However, like our survey, both these surveys and the full-text corpus use convenience samples. Given the method of sampling that was used it is unlikely that these samples are representative as certain demographics may be more likely to participate in the survey than others. Because we have no clear hypotheses or information about what the potential sources of bias may be, we did not include them in the statistical models below. We do urge readers to interpret the results of our analyses with caution, as they may be biased by our method of sampling.

Disagreement

The responses (agree or disagree) for all conditions for all cases are summarized in figure 2. For each condition, the participants were more likely to accept the abstract for a conference than they

were to agree with the ranking decision (see supplemental figure S4.4). This suggests that respondents did not interpret the agree question as one about scientific quality and that the abstracts were generally seen as academically acceptable.

The estimated proportion of 'agree' for each condition for each case is listed in table 5. Disagreement (the size of the minority opinion) within conditions was above 0.25 for the entire 0.80 hdi for 4 out of 10 conditions (combining the three cases). Thus, our hypothesis that there would be strong disagreement about grey-area cases was not confirmed according to the criteria we had selected. This was due in particular to the plant case, which had relatively high levels of agreement. Still, all conditions showed at least moderate disagreement, with an average proportion of 27.84% for the minority opinion across all conditions (see supplemental figure S4.2), and disagreement means ranging between 17.7% and 45.9%.

Of course, these results should be interpreted with caution. On the one hand, they might underestimate true rates of disagreement if the plant case was too clear-cut and did not represent a true grey-area case. More generally, there were substantial differences in disagreement between conditions and cases, with high levels of agreement for all three plant conditions and the mtDNA condition of the frog case. This suggests that levels of disagreement may be very case-dependent, and it remains an open question to what extent we can generalize findings about them to other cases.

On the other hand, this study might overestimate disagreement as well. First, the vignettes were explicitly designed to be grey-area cases that are likely to elicit disagreement. This means that the results only apply to such cases, and not across the whole hierarchical realm covered by taxonomy (in line with findings by Faurby et al. 2016). Many cases of species delimitation will be uncontentious. Second, the vignettes in this study were short abstracts, and participants were asked to evaluate the abstracts even if they were outside their taxon of specialization. This is unlike taxonomic reality, in which ranking decisions are typically not made based on information that can be given in an abstract of 150 words, and taxonomists rarely have to make ranking decisions outside their taxon of expertise.

Thus, it may also be that part of the disagreement in these cases was caused by the lack of information in the abstracts. This is suggested by the fact that in the frog case, disagreement decreases with more information (i.e. going from neutral to one of the conditions with extra evidence). However, while it is true that the vignettes provided less information than taxonomists typically work with, it should be remembered that working with little information is the sad reality of many taxonomic decisions. Thus, while the lack of information might exacerbate the disagreement here, we do not think it is merely a product of our methods.

Drivers of disagreement: species concepts and operationalization

Supplemental tables S4.6, S4.7 and S4.8 list the coefficients and hdi's for the selected variables of interest for models 1 – 3 (full results in S3, and not reported here to avoid the 'table 2 fallacy' (Westreich and Greenland 2013)). In all three models, the influence of species concepts on accepting the species descriptions was close to zero. As expected, the effect of species concepts was strongest in the flatworm case, which was centered on gene flow (fig 3). Because reproductive isolation is the main criterion for species status in the BSC, we expected that the difference in expected proportions of 'agree' for the 'geneflow' and 'non-geneflow' conditions would be largest for proponents of the BSC: they should accept the species if there is no geneflow, and reject it if there is geneflow. However, not only was there a substantial group of proponents of the BSC that accepted the species even under the 'gene flow' condition (mean expected proportion of 0.63), posterior predictive sampling from model 3 setting the entire population in turn to the various combinations of treatments and species concepts also showed that we should expect almost no difference between the different species concepts in how important gene flow is (fig. 4). That is, the difference in 'agree' between 'geneflow' and 'no geneflow' was nearly identical across species concepts. More generally, levels of disagreement within species concepts were very similar to levels of disagreement across species concepts (supplemental figure S4.3). All this suggests that the influence of species concepts

on ranking decisions was small, and, if there was any, not directly related to the content of the concepts.

Contrary to species concepts, operationalization did seem to have a strong influence on agreeing with the ranking decision in the abstract. The model for the frog case, which was designed to test the influence of operationalization, shows that rates of disagreement differed substantially between treatments. In particular, evidence of mtDNA differentiation appeared to be a far stronger reason to recognize the frog as a species than morphological and ecological evidence. Figure 5 shows that while posterior predictive proportions of 'agree' hardly differ between species concepts, they differ strongly between morphology on the one hand, and mtDNA and habitat on the other hand. This shows that for the frog abstract, operationalization was far more influential than species concepts. The difference between morphological evidence and the other operationalizations also differed between groups, with taxonomists working in low-income countries accepting it more often as sufficient for species status (figure 6). We suspect this may be the case because taxonomists in low-income countries do not always have resources to produce molecular evidence and therefore have to rely on morphological evidence more often and because the tradition of morphology-based taxonomy is therefore particularly strong in low-income countries.

Drivers of disagreement: conservation values

We found no difference (mean: 0.003, 80%hdi: -0.061, 0.057; see supplemental figure S4.9) in the estimated proportion of 'agree' between the 'threatened' and 'abundant' version of the plant case for the sample of the hypothesis tests. This suggests that in this case, conservation status did not influence ranking decisions, and that our second hypothesis, concerning a role for conservation values in taxonomic decision-making, is therefore disconfirmed. This goes against the commonly made claim in the literature that taxonomists sometimes tend to recognize threatened groups as species merely to improve their chances of getting funding for conservation action (Isaac et al. 2004, Conix

2019). It should be noted, however, that there was less disagreement in general about the plant case than about the other two cases. Thus, as already mentioned, an alternative explanation may be that the vignette was not considered as grey-area case by the respondents, and because of that did not show conservation values to play a role.

It should also be noted that for the extended sample, with additional sampling effort, model 1 expects the proportion of 'agree' to be higher for groups with the 'threatened' version of the vignette if we take posterior predictive samples from the model assuming the demographics of our study population. This is not due to differences in the coefficients for 'threatened' and 'abundant' (see supplemental figure S4.5), but due to the fact that the model finds a clear difference between 'threatened' and 'abundant' for taxonomists working in low-income countries (which we tried to sample from in the second round of sampling, excluded from the hypothesis test) (see fig. 7). We speculate that this may be the case because the fictional plant case was set in a tropical environment. Taxonomists from low-income countries are more likely to be active in such environments on a daily basis, and may thus have felt a stronger connection to the plant group in question. Another possibility is that taxonomists from the global south are more concretely aware of the ongoing extinction crisis because the tropics are such a major stronghold of the world's biodiversity.

It is important to highlight the limitations of the plant case as a test of the role of non-taxonomic values in ranking decisions. For one, we only looked at conservation values. Other sociological factors may well play a role that may not have been captured by the plant case. For example, in all three models the taxon of specialization as well as the income status of the country of activity seemed to affect the tendency to agree with the ranking decision, with insect taxonomists and taxonomists from low-income countries showing a tendency to split. Thus, it may well be that sociological factors such as country of residence and training, varying academic culture and traditions in different taxonomic communities are non-taxonomic considerations that influence ranking decisions. However, even if the causal model we assume (figure 1) implies that the coefficients for

these variables are meaningful as an indication of direct (not total) causal effect, it is not possible with our data to draw definitive conclusions about this, and hypothesis-driven follow-up research is needed to confirm and flesh out these patterns.

Conclusions and prospects

This survey indicates that there was at least moderate taxonomic disagreement about the fictional grey-area cases we presented. Even though taxonomists were given the same information about a group of organisms, there was an average 28% disagreement about the groups' status as a separate species. Equally important is the light that our survey sheds on the drivers of this disagreement. Unlike what many researchers seem to believe, differences in adherence to species concept do not appear to lead to more differences in the observed taxonomic decisions, and adherence to the same species concept does not lead to lower levels of disagreement. The concrete operationalization of species concepts seems far more important for explaining taxonomic disagreement. Since these operationalizations are not strictly tied to a single species concept, this indicates that the disagreement may often be practical (information and methods) rather than theoretical (concepts). Lastly, contrary to our expectations, conservation values did not seem to motivate taxonomic decisions, at least not in general. Again, this contrasts with the attention that is given to the role of values for taxonomic decision-making in both the philosophical and biological literature (Isaac et al. 2004, Ludwig 2016, Conix 2019).

We draw two main concrete conclusions from this. First, our results suggest that, at least in grey-area cases, some degree of subjectiveness is sometimes probably hard to avoid if we insist on using the current Linnaean system where taxa are given specific ranks: while disagreement was not as high as we expected, there was at least moderate disagreement about every case. The fact that disagreement is probably most common in grey-area cases should not be taken to entail that more information on the groups will (always) solve disputes on species status. Although we did find, in the

frog case, that more information might sometimes reduce disagreement, this is not the silver bullet some consider it to be. For one, evidence is sometimes lacking and it is not always possible or feasible to gather more evidence. Moreover, there are also cases in which different lines of evidence or methods conflict (Satler et al. 2013). Simply adding data is unlikely to solve all problems because speciation inevitably is a multifaceted and gradual process. Collecting more data does not turn shallow divergence into deep divergence, and even with an abundance of data (dichotomous) decisions remain difficult in such cases.

This reality need not reflect negatively on taxonomy as a scientific discipline, or on the work of taxonomists – there are parallels in other, equally respectable fields of science (Slater 2017, Cuypers and De Block 2023). In genuine grey-area cases, the uncertainty about species status and the resulting taxonomic disagreement refer to the ranking part of taxonomy and do not necessarily reflect ignorance about what a species is or about particular characteristics of the group under consideration. Rather, they are an inevitable consequence of the application of a binary system onto a non-binary, continuous reality (Zachos et al. 2020, Thiele et al. 2021).

We do believe, however, that taxonomists should keep this reality of subjectivity in mind, and should take some measures to alleviate unwanted consequences. For example, we believe it is important that taxonomists provide full transparency on why they decide what they decide. Taxonomists should provide detailed methodological information – as our results show, operational choices matter – and information on how they interpret their results and translate them into taxonomic decisions. One step in this direction could be to register taxonomic methods and criteria for attributing species status in advance. As some of the authors of this study have argued elsewhere, the preregistration of research methods has beneficial effects on transparency and clarity in many disciplines, and could also be of use in taxonomy (Conix et al. 2023).

The degree of subjectiveness involved in taxonomic decision-making in grey-area cases should also be acknowledged when assessing – at times vehement – taxonomic disagreements. If

disagreements concern empirical questions, for example on evolutionary patterns in the groups under considerations, they obviously have scientific value. But if disagreements turn out to be a pure matter of appreciation, it may not always be useful to pursue debates about them endlessly, given urgent demands for clear and stable taxonomies in biology and beyond. Rather, it may be advisable to take recourse to procedures suited to arbitrate 'executive' issues of that kind, for example through some form of taxonomic list governance (Garnett et al. 2020). This is precisely what the four main global bird lists are currently doing, unifying their diverging lists through a voting procedure (McClure et al. 2020, Cuypers and De Block 2023).

The second main implication of our results is that a shift may be needed in what philosophers and biologists should focus on when they study the conceptual side of the 'species problem'. Our results suggest that the research community should probably spend more time researching the role of operationalization in ranking decisions, and focus less on studying how species concepts and non-epistemic values may shape taxonomy. This dovetails nicely with the first implication, as what we need in particular is renewed reflection on how to deal with grey-area cases in taxonomic practice.

Acknowledgements

Stijn Conix' work for this paper was funded by the Fonds de la Recherche Scientifique - FNRS under grant no. T.0177.21. Vincent Cuypers' work for this paper was funded by the Research Council Flanders (FWO) under grant no. 3H200026.

References

Agapow P-M, Bininda-Emonds ORP, Crandall KA, Gittleman JL, Mace GM, Marshall JC, Purvis A. 2004. The Impact of Species Concept on Biodiversity Studies. The Quarterly Review of Biology 79: 161–179.

Barr L. 2023. CausalGraphicalModels.

Camargo A, Sites JJ. 2013. Species Delimitation: A Decade After the Renaissance. The Species Problem - Ongoing Issues.

Capretto T, Piho C, Kumar R, Westfall J, Yarkoni T, Martin OA. 2022. Bambi: A Simple Interface for Fitting Bayesian Linear Models in Python. Journal of Statistical Software 103: 1–29.

Cinelli C, Forney A, Pearl J. 2022. A Crash Course in Good and Bad Controls. Sociological Methods & Research 00491241221099552.

Conix S. 2018. Integrative taxonomy and the operationalization of evolutionary independence. European Journal for Philosophy of Science 8: 587–603.

Conix S. 2019. Taxonomy and conservation science: interdependent and value-laden. History and Philosophy of the Life Sciences 41: 15.

Conix S, Block A de, Cuypers V, Zachos F. 2022. Exploring the reasons of diverging views in taxonomy.

Conix S, Cuypers V, Zachos F, Artois T, Monnens M. 2023. A plea for preregistration in taxonomy. Megataxa 10: 1–14.

Cotterill FPD, Taylor PJ, Gippoliti S, Bishop JM, Groves CP. 2014. Why one century of phenetics is enough: response to "Are there really twice as many bovid species as we thought?" Systematic Biology syu003.

Cuypers V, De Block A. 2023. Resolving Conceptual Conflicts through Voting. Foundations of Science.

Faurby S, Eiserhardt WL, Svenning J-C. 2016. Strong effects of variation in taxonomic opinion on diversification analyses. Methods in Ecology and Evolution 7: 4–13.

Garnett ST, Christidis L, Conix S, Costello MJ, Zachos FE, Bánki OS, Bao Y, Barik SK, Buckeridge JS, Hobern D, Lien A, Montgomery N, Nikolaeva S, Pyle RL, Thomson SA, Dijk PP van, Whalen A, Zhang Z-Q, Thiele KR. 2020. Principles for creating a single authoritative list of the world's species. PLOS Biology 18: e3000736.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del Río JF, Wiebe M, Peterson P, Gérard-Marchant P, Sheppard K, Reddy T, Weckesser W, Abbasi H, Gohlke C, Oliphant TE. 2020. Array programming with NumPy. Nature 585: 357–362.

Harris J, Froufe E. 2005. Taxonomic inflation: Species concept or historical geopolitical bias? Trends in Ecology & Evolution 20: 6–7.

Heller R, Frandsen P, Lorenzen ED, Siegismund HR. 2013. Are there really twice as many bovid species as we thought? Systematic Biology 62: 490–493.

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering 9: 90–95.

Isaac NJB, Mallet J, Mace GM. 2004. Taxonomic inflation: Its influence on macroecology and conservation. Trends in Ecology & Evolution 19: 464–469.

Karl SA, Bowen BW. 1999. Evolutionary significant units versus geopolitical taxonomy: Molecular systematics of an endangered sea turtle (genus Chelonia). Conservation Biology 13: 990–999.

Kumar R, Carroll C, Hartikainen A, Martin O. 2019. ArviZ a unified library for exploratory analysis of Bayesian models in Python. Journal of Open Source Software 4: 1143.

Ludwig D. 2016. Ontological choices and the value-free ideal. Erkenntnis 6: 1253–1272.

McClure CJW, Lepage D, Dunn L, Anderson DL, Schulwitz SE, Camacho L, Robinson BW, Christidis L, Schulenberg TS, Iliff MJ, Rasmussen PC, Johnson J. 2020. Towards reconciliation of the four world bird lists: hotspots of disagreement in taxonomy of raptors. Proceedings of the Royal Society B: Biological Sciences 287: 20200683.

McKinney W. 2010. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference 56–61.

Mishler BD, Wilkins JS. 2018. The Hunting of the SNaRC: A Snarky Solution to the Species Problem. Philosophy, Theory, and Practice in Biology 10.

Neate-Clegg MHC, Blount JD, Şekercioğlu ÇH. 2021. Ecological and biogeographical predictors of taxonomic discord across the world's birds. Global Ecology and Biogeography 30: 1258–1270.

Pence CH, Conix S. Forthcoming. Mapping Controversy: A Cartography of Taxonomy and Biodiversity for the Philosophy of Biology. in Ulatowski J, Weijers D, and Systma J, eds. Advances in Experimental Philosophy and Corpus Methods. Bloomsbury.

Pušić B, Gregorić P, Franjević D. 2017. What do Biologists Make of the Species Problem? Acta Biotheoretica 65: 179–209.

van Ravenzwaaij D, Cassey P, Brown SD. 2018. A simple introduction to Markov Chain Monte–Carlo sampling. Psychonomic Bulletin & Review 25: 143–154.

Salvador RB, Cavallari DC, Rands D, Tomotani BM. 2022. Publication practice in Taxonomy: Global inequalities and potential bias against negative results. PLOS ONE 17: e0269246.

Salvatier J, Wiecki TV, Fonnesbeck C. 2016. Probabilistic programming in Python using PyMC3. PeerJ Computer Science 2: e55.

Sangster G. 2014. The application of species criteria in avian taxonomy and its implications for the debate over species concepts: Application of species criteria in practice. Biological Reviews 89: 199–214.

Satler JD, Carstens BC, Hedin M. 2013. Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). Systematic Biology 62: 805–823.

Slater MH. 2017. Pluto and the platypus: An odd ball and an odd duck - On classificatory norms. Studies in History and Philosophy of Science Part A 61: 1–10.

Stankowski S, Ravinet M. 2021. Quantifying the use of species concepts. Current Biology 31: R428–R429.

Tattersall I. 2007. Madagascar's Lemurs: Cryptic diversity or taxonomic inflation? Evolutionary Anthropology: Issues, News, and Reviews 16: 12–23.

Thiele KR, Conix S, Pyle RL, Barik SK, Christidis L, Costello MJ, van Dijk PP, Kirk P, Lien A, Thomson SA, Zachos FE, Zhang Z-Q, Garnett ST. 2021. Towards a global list of accepted species I. Why taxonomists sometimes disagree, and why this matters. Organisms Diversity & Evolution.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17: 261–272.

Waskom M, Gelbart M, Botvinnik O, Ostblom J, Hobson P, Lukauskas S, Gemperline DC, Augspurger T, Halchenko Y, Warmenhoven J, Cole JB, Hoeven E ter, Ruiter J de, Vanderplas J, Hoyer S, Pye C, Miles A, Swain C, Meyer K, Martin M, Bachant P, Molin S, Quintero E, Kunter G, Villalba S, Brian, Fitzgerald C, Evans C, Williams ML, O'Kane D. 2022. mwaskom/seaborn: v0.12.2 (December 2022).

Westreich D, Greenland S. 2013. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. American Journal of Epidemiology 177: 292–298.

Willis SC. 2017. One species or four? Yes!...and, no. Or, arbitrary assignment of lineages to species obscures the diversification processes of Neotropical fishes. PLOS ONE 12: e0172349.

World Bank. 2022. Low income | Data. (13 February 2023; https://data.worldbank.org/country/XM).

Zachos FE. 2022. Critique of Taxonomic Reason(ing): Nature's Joints in Light of an 'Honest' Species Concept and Kurt Hübner's Historistic Philosophy of Science. Species Problems and Beyond. CRC Press.

Zachos FE, Christidis L, Garnett ST. 2020. Mammalian species and the twofold nature of taxonomy: a comment on Taylor et al. 2019. Mammalia 84: 1–5.

Author Biographies

Stijn Conix (stijn.conix@uclouvain.be) is a philosopher of science at the Université catholique de louvain , Belgium. He uses empirical methods to answer conceptual questions in taxonomy and biology more generally .

Vincent Cuypers (vincent.cuypers@uhasselt.be) is a doctoral student in biology and philosophy of science at Hasselt University and KU Leuven, Belgium, working at the interface of science studies and biological practice. His research interests include taxonomy, general terrestrial and marine ecology, farmland ecology, science studies and metaphilosophy.

Frank E. Zachos (frank.zachos@NHM-WIEN.AC.AT) is an evolutionary zoologist with the Natural History Museum Vienna, Austria, and the Department of Genetics at the University of the Free State in Bloemfontein, South Africa, who has a longstanding interest in both the philosophy and practice of taxonomy and phylogenetics.

Andreas De Block (andreas.deblock@kuleuven.be) is a philosopher of science at the KU Leuven, Belgium. His main interests lie in philosophy of medicine and experimental philosophy.

Table 1: Overview of the different cases and vignettes with the number of respondents for each condition for each case. Because we collected additional data after the preregistered period, the number of participants for the hypothesis tests differs from the total number of participants.

Case	Condition	N (total)	N (hypothesis test)
Plant	Neutral	143	127
	Threatened	151	134
	Abundant	143	128
Frog	Neutral	119	103
	Morphology	107	98
	mtDNA	105	90
	Ecology	110	100
Flatworm	Neutral	155	134
	Gene flow	139	128
	No gene flow	141	124

Table 2: Statistical models for the exploratory analysis of ranking decisions

	Case	Outcome	Cause of interest	Implementation cause	Controled for
Model 1	Plant	Agree	Values	Species concept; Treatment (neutral, threatened, abundant)	Income, taxon of specialization.
Model 2	Frog	Agree	Operationalization	Species concept ; Treatment (neutral, morphology, DNA, habitat)	Income, taxon of specialization.
Model 3	Flatworm	Agree	Species concept	Species concept; Treatment (gene flow, no gene flow, neutral)	Income, taxon of specialization.

Table 3: Years since start of taxonomic activity, continent of residence and preferred species concept of the respondents in the complete sample.

	Continent of residence						Species Concept					
	Africa (n=17)	Asia (n=50)	Europe (n=191)	North America (n=100)	Oceania (n=27)	South America (n=56)	BSC (n=104)	ESC (n=103)	GCSC (n=12)	dPSC (n=120)	mPSC (n=42)	other (n=45)
0 to 5 (%)	17.6	8	8.4	5	3.7	5.4	5.8	10.7	0	5.8	9.5	6.7
6 to 10 (%)	17.6	22	13.6	6	3.7	12.5	11.5	13.6	0	10.8	14.3	13.3
11 to 20 (%)	23.5	44	26.2	17	7.4	30.4	18.3	33	0	31.7	23.8	15.6
21 to 30 (%)	5.9	12	26.2	25	14.8	21.4	25	19.4	41.7	20.8	11.9	28.9
31+ (%)	35.3	14	25.7	47	70.4	30.4	39.4	23.3	58.3	30.8	40.5	35.6

Table 4: Taxon specializations of respondents in the sample separated by whether the respondent is active in a high or low income country.

	Low income (%)	High income (%)
Algae (n=6)	16.7	83.3
Birds (n=11)	36.4	54.5
Fishes (n=28)	14.3	85.7
Fungi (n=5)	0	100
Insects (n=135)	34.1	65.9
Mammals (n=30)	36.7	63.3
Molluscs (n=19)	21.1	73.7
Non-insect arthropods (n=61)	32.8	65.6
Non-vertebrate deuterostomes (n=4)	50	50
Plants (n=59)	32.2	67.8
Prokaryotes (n=3)	0	100
Protists (non-algae) (n=7)	42.9	57.1
Remaining invertebrates (n=46)	26.1	71.7
Reptiles and Amphibia (n=33)	45.5	54.5

Table 5: Estimated proportion of 'agree' for each condition for each of the three cases. Columns starting with 'Hyp' summarize results from the hypothesis test model. Columns starting with 'FD' give the marginal at the mean for the treatments obtained from models 1, 2 and 3.

Case	Condition	Hyp Mean	Hyp sd	Hyp 10%hdi	Hyp 90%hdi	Hyp Minority > 0.25	FD mean	FD 10%hdi	FD 90%hdi
Plant	Neutral	0.177	0.033	0.131	0.216	no	0.135	0.084	0.182
	Abundant	0.221	0.036	0.173	0.265	no	0.162	0.101	0.212
	Threatened	0.219	0.035	0.17	0.261	no	0.232	0.167	0.292
Frog	Neutral	0.459	0.048	0.395	0.518	yes	0.450	0.380	0.524
	DNA	0.82	0.04	0.772	0.874	no	0.863	0.818	0.916
	Habitat	0.731	0.043	0.677	0.788	no	0.701	0.634	0.774
	Morphology	0.609	0.048	0.548	0.672	yes	0.611	0.538	0.686
Flatworm	Neutral	0.696	0.039	0.648	0.748	yes	0.666	0.605	0.736
	Gene flow	0.659	0.041	0.608	0.713	yes	0.675	0.608	0.745
	No gene flow	0.758	0.038	0.712	0.809	no	0.765	0.712	0.826

Figure 1: Causal model of ranking decisions. 'Agree' is the outcome, 'Treatment' are the various versions of the vignettes for each of the three cases. The figure was made using the causalgraphicalmodels (Barr 2023) package.

Figure 2: Responses to the question 'Do you agree with the ranking decision in the abstract' for each of the conditions of each of the cases for the full dataset. 'Agree' indicates agreement with recognizing a new species rather than lumping it.

Figure 3: Density plot of the difference between the expected proportion of 'agree' for the BSC and for other species concepts using posterior predictive samples. These posterior predictive samples were drawn from model 3, leaving the demographic characteristics of the sample intact but changing the species concept to each of the included concepts for the entire sample.

Figure 4: Density plots of how the BSC and each of the other species concepts differ in the difference of proportion 'gene flow' and 'no gene flow' in posterior predictive samples drawn from model 3 (flatworm case; drawn for each combination of treatments and species concepts, keeping the other demographic properties of the sample intact).

Figure 5: Density plots of the expected differences in expected proportion of 'agree' between the BSC and other concepts (all in blue and solid), and between morphology and other treatments (all in red and dashed). The expected proportions were generated using posterior predictive samples from model 2, keeping the other demographic properties of the sample intact.

Figure 6: Density plots of the expected differences in expected proportion of 'agree' between high- and low-income countries for the difference between neutral and other treatments. The expected proportions were generated using posterior predictive samples from model 2, keeping the other demographic properties of the sample intact.

Figure 7: Difference in proportion of 'agree' between 'neutral' and the two other treatments for posterior predictive samples from model 1 with the entire sample set to 'high income' (top) and 'low income' (bottom). This clearly shows that participants working in low-income countries were more likely to agree with the ranking decision in the 'threatened' condition.

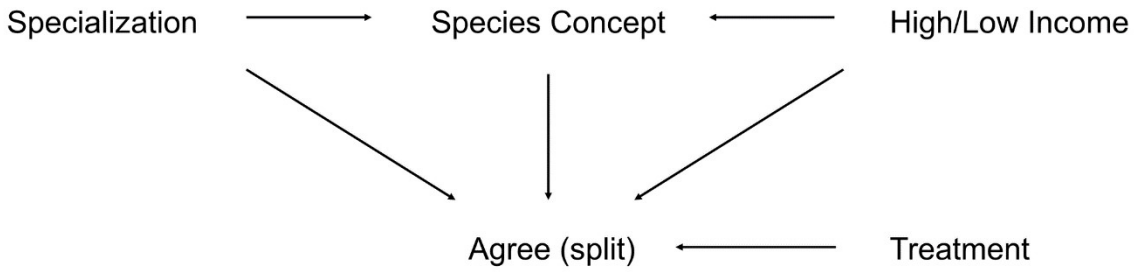

Fig. 1

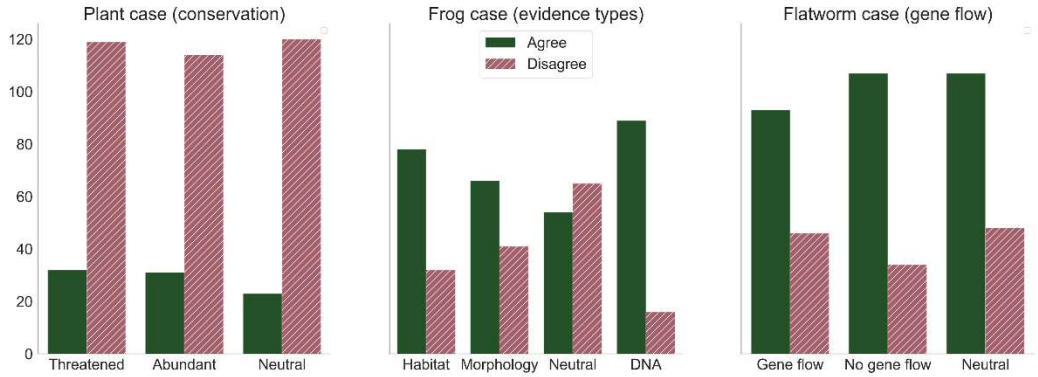

Fig. 2

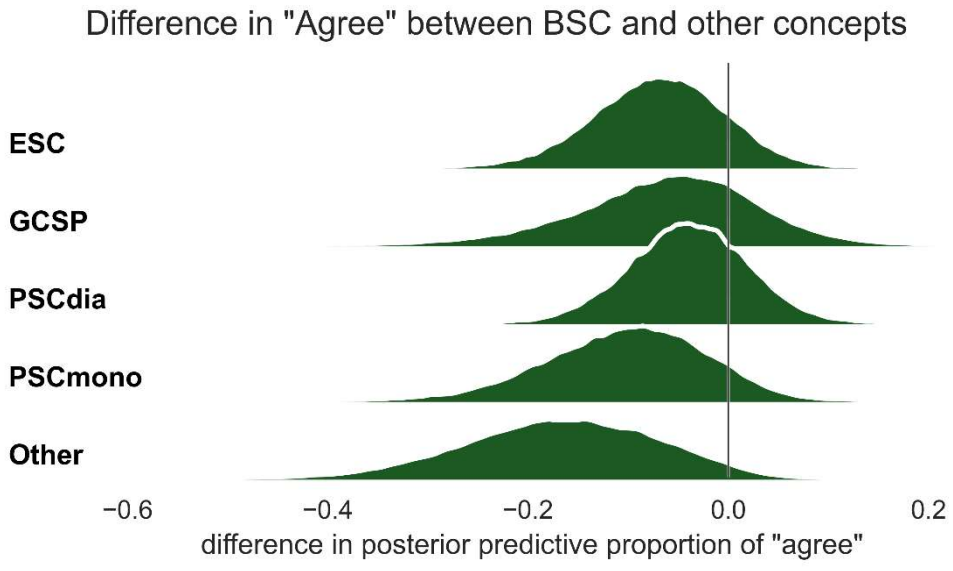

Fig. 3

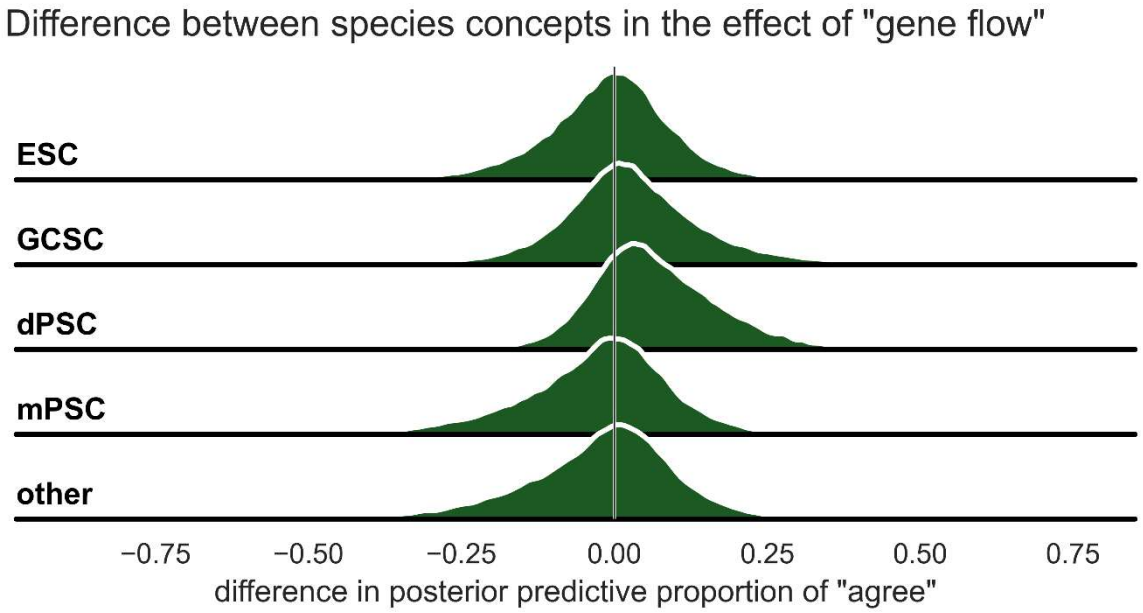

Fig. 4

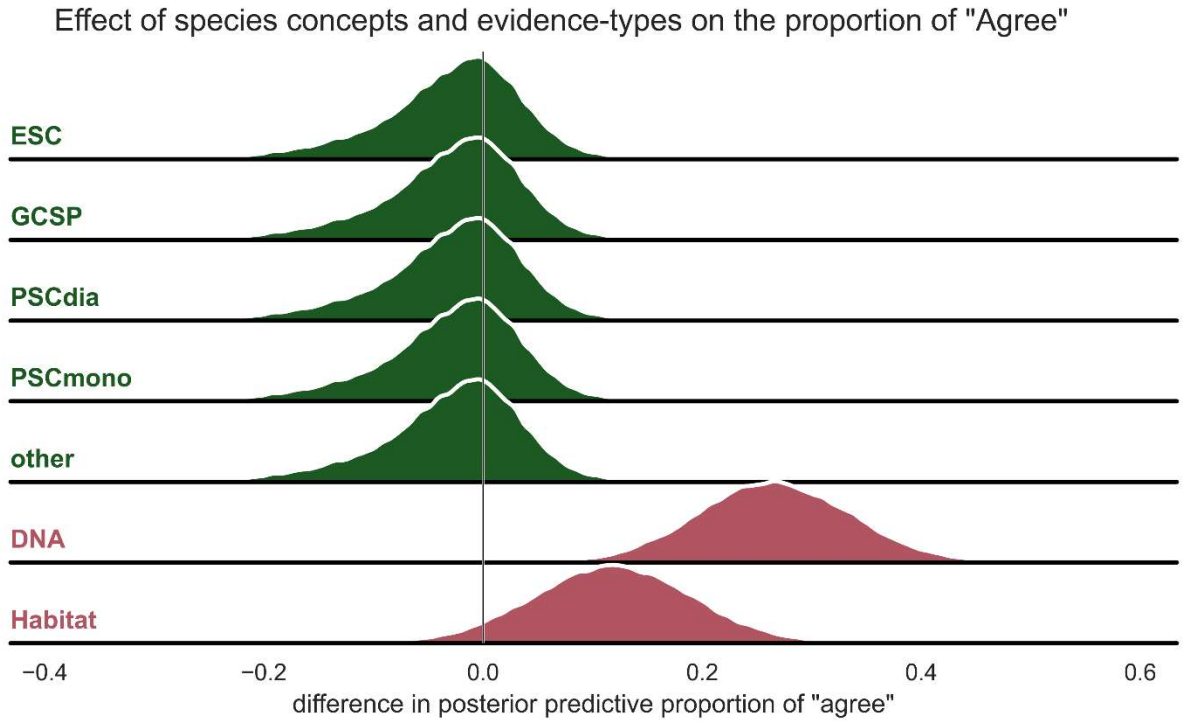

Fig. 5

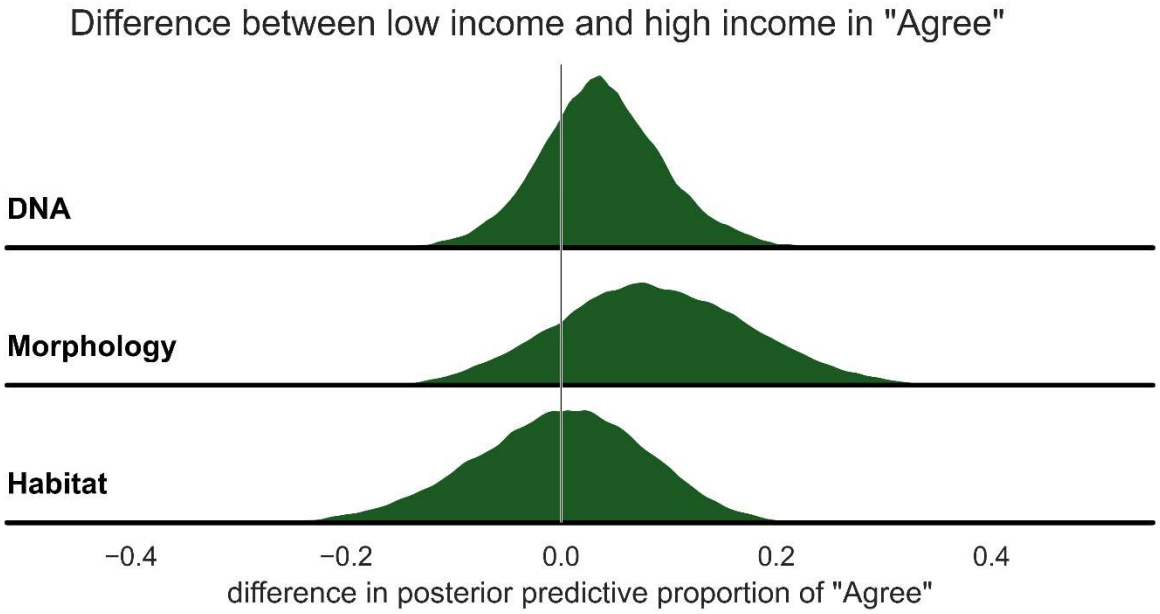

Fig. 6

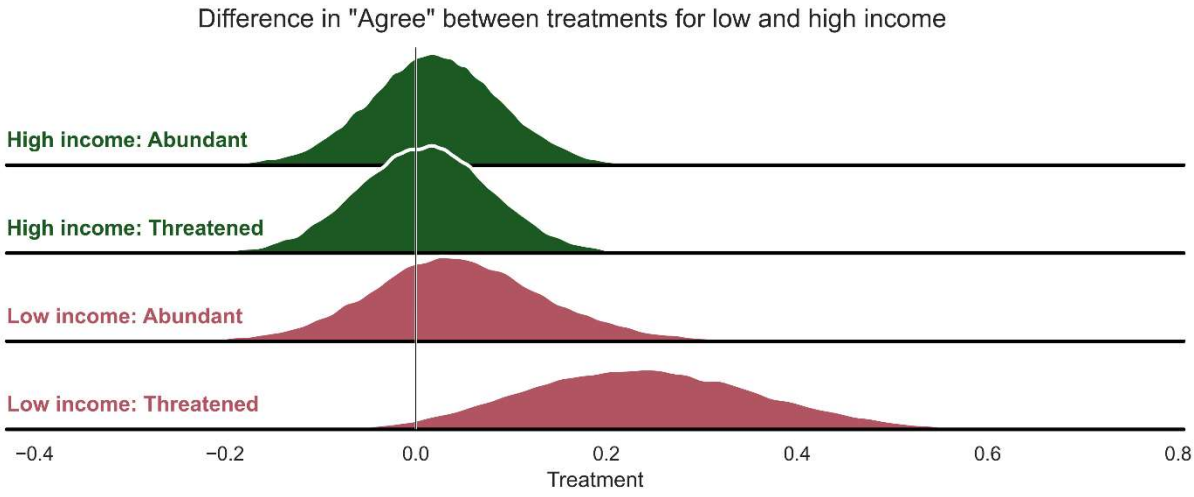

Fig. 7