Technische Universiteit Eindhoven

Beyond Binary: Segmentation-Based Health Identifier for Coral Reefs

Social–Legal–Ethical (SLE) Essay

How should responsibility for decisions informed by an AI coral-bleaching classifier be allocated among stakeholders?

David Mandado, Marc Boglioni, David Ilisei, Pelayo Davara, Timo Wijnen

Group 13 — Capstone Data Challenge

Supervising Tutor: Michiel Cevaal

Quartile 1, 2025

Abstract

Coral reefs face accelerating degradation due to serious bleaching events caused by rising sea temperatures. Artificial intelligence offers new possibilities for monitoring reef health at scale, but the delegation of environmental decisions to algorithmic systems raises complex questions of responsibility. This essay addresses how responsibility for decisions informed by an AI coral-bleaching segmentation model should be distributed among stakeholders, including data scientists, implementing agencies, and local communities. Drawing on case studies from NOAA and the Great Barrier Reef Marine Park Authority, the paper analyses three dimensions of responsibility: social, legal, and ethical. The social dimension considers how algorithmic classifications affect tourism, fisheries, and local livelihoods; the legal dimension identifies where accountability resides under existing environmental governance frameworks; and the ethical dimension addresses fairness and bias in model predictions. We argue that while operational responsibility remains with environmental agencies, developers and data providers share a duty of transparency, bias auditing, and public communication to ensure decisions remain justifiable and proportionate.

1 Introduction

Coral reefs are among the most biodiverse ecosystems on the planet and are essential to food security, coastal protection, and tourism livelihoods. However, they are increasingly threatened by coral bleaching events linked to ocean warming and pollution. Monitoring bleaching at the scale required for effective response is expensive and labor-intensive. As a result, governments and NGOs have begun to adopt AI-driven systems that automatically classify and quantify coral health from imagery.

While such systems promise efficiency and consistency, they also raise questions about how algorithmic assessments shape real-world decisions (such as reef closures, funding allocation, and fishing restrictions) and who bears responsibility if those decisions are wrong. This essay uses the example of a segmentation-based coral-health model developed with Reef Support to explore these questions. The segmentation model looks into different features of coral color (luminance, saturation, raw-red light and albedo) and texture (glcm correlation, Laplacian variance and local binary pattern).

Our central research question is:

How should responsibility for decisions informed by an AI coral-bleaching classifier be allocated among stakeholders?

Now, to fully grasp the importance of this question, it is crucial to understand all the different parties that might be involved in coral reef maintenance and the consequences that decisions taken upon the information provided by a bleaching classifier have. In this essay we will focus on three different parties: tourism, fisheries and government.

Tourism

To begin with, the decisions made based on the model's output can influence tourism in a specific area. If such an area is determined to be "at-risk," restrictions might be applied to the number of visitors, or it might even be closed for tourism altogether. This could lead to an economic and social impact on that area. As a matter of fact, the Great Barrier Reef is the perfect example. The GBR is estimated to contribute A$6.4 billion nationally and is linked to 64.000 jobs (O'mahoney et al., 2017), so even the slightest downturn has a major impact. Actually, during the bleaching of 2016-2017, at Cairns, domestic visitor numbers and revenue from interstate tourism were down (Smee, 2018). In another area of Australia, Whitsundays, tourist figures were down 50% and it was said to be as bad as it was during the global financial crisis (Rebgetz, 2017). A report from the Australia Institute predicted the social impacts the coral bleaching of 2016-2017 would have on the Reef's tourism. They stated that the Reef's tourism areas were at risk of losing over one million visitors per year. Along with visitor numbers, the potential loss of tourism revenue represents almost one-third of the $3.3 billion spent by holiday visitors to Reef regions each year, which supports between 39,000 and 45,000 jobs. Around 10,000 jobs were expected to be at risk from decreased visitation and spending if severe coral bleaching of the Reef continued (Swann et al., 2016).

Fisheries

Additionally, if the segmentation model flags a site as severely bleached, it may lead to temporary fishing limits to speed up reef recovery. For example, prohibiting the fishing of herbivores (like parrotfish or surgeonfish) is a common practice to support coral reefs. Many studies have found that herbivores provide resilience to reefs because they keep macroalgae at low levels after a disturbance and allow corals to recover (Burkepile & Hay, 2008). As NOAA stated, herbivore management plays a role in a broader strategy to manage and reduce threats to coral reefs (Fisheries, 2019). In fact, part of the Hawaii recovery plan after the 2015 mass bleaching consisted of establishing a network of permanent no-take Marine Protected Areas (MPAs) and establishing a network of Herbivore Fishery Management Areas (HFMAs) (Rosinski et al., 2017). These restrictions may influence specific fisheries whose revenue depends on the commercialization of herbivores.

Government spending

Flagging an area as severe bleaching is also a burden for the government of that area. Implementing diverse measures for coral support increases government spending, whether through monitoring, protection, or restoration efforts, all of which require more public spending. Back in 2023, during the Florida Keys heatwave, NOAA set up temporary Special Use Areas to protect relocated nursery corals, meaning that the entry was restricted and time-limited until the temperature

dropped and corals could be relocated back (Atwell, 2023). To implement these measures, government spending is needed. As a matter of fact, each year, NOAA's Coral Reef Conservation Program is awarded over $8 million in grants and cooperative agreements (US Department of Commerce & Administration, n.d.). Having analyzed how all of these parties have been affected and react to coral reef maintenance due to coral bleaching, we hope the reader understands the importance the research question this essay faces.

2 Sub-question A

As seen before, the decisions made based of an ai decision tool have rather an important social impact, so it is important to look into the where and why a segmentation model might have erroneous outputs. This is why we came up with a sub question:
How should responsibility for decisions informed by an AI coral-bleaching classifier be allocated among stakeholders?

For this situation, the most probable type of bias a model will face is dataset shifting. Dataset shift is a common problem. It happens when the joint distribution of inputs and outputs differs between training and test stages (Quiñonero-Candela et al., 2008). In this concrete scenario, dataset shifting might happen if the model was trained with images from Site A (green water and bright sand) and is later tested with coral images of Site B (blue water and darker corals). This bias in a model can lead to the model becoming an expert on how to detect coral bleaching for certain conditions, for example Site A, but as a consequence of this, when presented with a coral image living under different conditions, it overrates or underrates the severity of bleaching. In other words, the model is site-specific, which leads to the presence of false positives and false negatives. A false positive would be when the model identifies high bleaching for Site A when in reality is not severe. On the other hand, a false negative would be when the model identifies no severe bleaching for Site B when it is in fact severely bleached. There are different measures that can be taken to prevent bias due to site specific conditions. At the beginning of the introduction of this essay, it was mentioned that our segmentation model looked into different features of each coral for both the color and texture of said coral. Now, to prevent any form of site specific bias, the contribution each feature has into coral bleaching is accounted for by making an educated guess of the weight of each feature using the following regression equation: $\beta = R_{xx}^{-1} r_{xy}$.

This equation takes into account the correlation between each feature and the percentage of area of coral bleached as well as the correlation between features, therefore not only having into account the direct effect of each feature to the bleaching of coral, but it also takes into account whether or not their effect on coral bleaching is not influenced as well by another feature. Since this equation was derived from standardized variables (Jöreskog, 1999) which meant that the feature scores needed to be standardized as well to achieve a higher accuracy of the weights. What standardizing the results means in terms of specific site bias, is that, not only do all the features now follow the same distribution, but they are all in the same scale as well. This is

very useful for the model to compare feature scores, and it prevents the model from learning that a certain feature is more relevant to coral bleaching as it has a higher score, therefore, it avoids scale privilege. By standardizing results, it supports transparent thresholds and consistent explanations to stakeholders. To show how this method reduces bias, the weights of each feature were calculated before and after standardizing the results. Here are the before and after weights for the color features:

	Before	**After**
Albedo	0.322	0.154
Luminance	0.135	0.198
Saturation	0.103	0.289
Raw red light	-0.139	-0.124

Table 1: Before vs. After metrics.

As it can be seen from the table, before standardizing the results, the model was learning that albedo had the highest percentage of importance towards coral bleaching, where in reality, saturation and luminance where the two features which have the most weight.

Not only where the scores standardized, but the median was taken into account instead of the mean. The way the features scores were calculated for an image, as it is a segmentation model, was calculating the score for each pixel in that image. Once this is done, to obtain an overall score for each feature, either the mean score or the median score can be computed. We calculated the correlation of color with the percentage of area bleached for both, and using the median scores we obtained a correlation of 0.58 for the mean scores and 0.51 for the median. However, we opted to go with the median score, as having into account the underwater, unclear images in different conditions, the median is more robust to outliers than the mean, therefore reducing the chance for the model to learn any site specific patterns.

The use of this equation however comes with different limitations. It assumes linear relationship between features and coral bleaching. In other words, it assumes that a change in coral whiteness or texture will have a steady and proportional effect on coral bleaching when it might not be de case. For example, a change in luminance might not have a major effect on bleaching at first but once a certain threshold is reached, bleaching can increase significantly with a minor change in luminance. Another limitation of using this regression equation is that is ignores non-linear effect. For example, it takes into account that the lower the luminance, the more bleached the coral is, or the lower the saturation, the higher the severity of bleaching. However, it might be the case that a coral is only bleached when there is low luminance and low saturation.

The use of this equation however comes with different limitations. It assumes linear relationship between features and coral bleaching. In other words, it assumes that a change in coral whiteness or texture will have a steady and proportional effect on coral bleaching when it might not be de case. For example, a change in luminance might not have a major effect on

bleaching at first but once a certain threshold is reached, bleaching can increase significantly with a minor change in luminance. Another limitation of using this regression equation is that is ignores non-linear effect. For example, it takes into account that the lower the luminance, the more bleached the coral is, or the lower the saturation, the higher the severity of bleaching. However, it might be the case that a coral is only bleached when there is low luminance and low saturation.

The methods explained above ensure that our model will either not be a victim of dataset bias or prevents it from having a significant effect to the final output.

3 Sub-question B

Why did we pick a simple, feature-based pipeline instead of a big black-box model—and why is that better for people who rely on these results?

We chose a feature-based approach because our data does not support a confident, fully supervised "bleaching severity" model, and pretending otherwise would create risks we can't justify. Ordinal health labels are scarce and inconsistent across sites, species, lighting and cameras. A large black-box model trained on that foundation might look accurate on paper while actually learning site quirks or color casts. That kind of hidden bias is hard to detect and even harder to explain.

Our pipeline limits machine learning to the part with clear ground truth, coral vs. non-coral segmentation, and then relies on simple, interpretable signals measured inside the masks: how pale the tissue looks (brightness, saturation, red channel) and how much fine texture is present (is the surface smooth or detailed). These choices are not just technical, they are ethical. People will use these outputs to make decisions that affect livelihoods and ecosystems. If a site is flagged as "at risk," local fishers, dive operators, and conservation staff deserve to know why. With our approach we can say, in plain terms, "this image is unusually pale and smoother than the site's norm," and show similar cases for context. That supports accountability and the right to question or correct the assessment.

There is also a fairness dimension. Reef programs differ widely in resources. A lightweight, explainable pipeline can run on modest hardware and be calibrated per site with a short expert session. That lowers costs, reduces the carbon footprint of monitoring, and makes the tool usable in places with limited connectivity. It also lowers the barrier to participation: local teams can understand, maintain, and adapt the system without relying on a distant model they can't inspect.

Finally, this design keeps uncertainty in view. Because the features are human-readable, we can communicate nuance (e.g., "paleness is high, texture change is moderate") and match actions to risk ("investigate" before "intervene"). The goal is not to replace judgment with a number, but to give communities a clear, honest signal grounded in what the data can reliably support. In that sense, choosing features over a black box is a social choice: it favors transparency over opacity, shared understanding over blind trust, and practical equity over sophistication for its own sake.

4 Conclusion

The deployment of an AI coral-bleaching segmentation model is a technological advance, but it also poses a governance challenge. As demonstrated above, algorithmic classifications can shape decisions that directly affect tourism, fisheries, and public spending. This makes it essential that the correct safeguards are in place. Socially, the consequences of mistakes made by the program can severely impact the communities that depend on the reefs for their livelihoods.

AI models like these should therefore function as decision-support tools, and not as final decision-makers. Clear documentation and known limitations help ensure that automated assessments remain accountable and scientifically grounded. Taking into account different coral conditions and calibration per site helps the model be a more precise decision-support tool . This can be seen throughout the use of standardized results for the feature scores, or the use of median calculation instead of mean, as well as site stratified splits. These are all methods used to prevent site specific, camera or condition bias, but have their own limitations.

Ultimately, responsible AI for reef management requires collaboration from data scientists, providers, policymakers, and local and global stakeholders . How can this collaboration take place? By combining all the information each can bring to the table. For example, thresholder actions. Instead of depending entirely on the models' output, make decisions based on external factors as well (e.g., NOAA Coral Reef Watch) to prevent spending on false alarms. For example, NOAA states that there is a risk of coral bleaching when the DHW value reaches 4 °C-weeks ("NOAA Coral Reef Watch Daily 5km Degree Heating Weeks", 2018). Then, actions by the government should be taken when the segmentation model identifies a coral as severely bleached and when the site DHW value reaches 4 °C-weeks.

References

Atwell, S. (2023, September). Noaa establishes temporary special use area to protect relocated florida keys coral nursery. Retrieved October 10, 2025, from https://sanctuaries.noaa.gov/news/sep23/coral-nursery-temporary-area.html

Burkepile, D. E., & Hay, M. E. (2008). Herbivore species richness and feeding complementarity affect community structure and function on a coral reef. *Proceedings of the National Academy of Sciences*, *105*(42), 16201–16206. https://doi.org/10.1073/pnas.0801946105

Fisheries, N. (2019, September). Restoring natural grazing processes can help coral reefs. Retrieved October 10, 2025, from https://www.fisheries.noaa.gov/feature-story/restoring-natural-grazing-processes-can-help-coral-reefs

Jöreskog, K. G. (1999, June). *How large can a standardized coefficient be?* (Technical Report) (Accessed on [insert access date]). LISREL Project / StatModel (Uppsala University). https://www.statmodel.com/download/Joreskog.pdf

Noaa coral reef watch daily 5km degree heating weeks. (2018). Retrieved October 10, 2025, from https://coralreefwatch.noaa.gov/product/5km/index_5km_dhw.php

O'mahoney, J., Simes, R., Redhill, D., Heaton, K., Atkinson, C., Hayward, E., & Nguyen, M. (2017). At what price? the economic, social and icon value of the great barrier reef.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2008). *Dataset shift in machine learning*. The MIT Press. https://doi.org/10.7551/mitpress/9780262170055.001.0001

Rebgetz, L. (2017). 'too big to fail': Great barrier reef valued at $56b. *ABC News*. Retrieved October 10, 2025, from https://www.abc.net.au/news/2017-06-26/great-barrier-reef-valued-56b-deloitte/8649936

Rosinski, A., Birkeland, C., Conklin, E., Williams, I., Gove, J., Gorospe, K., Oliver, T., & Walsh, W. (2017). *Identifying management responses to promote coral recovery in hawai'i coral bleaching recovery plan* (tech. rep.). University of Hawaii. https://dlnr.hawaii.gov/reefresponse/files/2016/09/CoralBleachingRecoveryPlan_final_newDARlogo.pdf

Smee, B. (2018). Domestic tourism to Great Barrier Reef falls in wake of coral bleaching. *The Guardian*. Retrieved October 10, 2025, from https://www.theguardian.com/environment/2018/jun/08/domestic-tourism-to-great-barrier-reef-falls-in-wake-of-coral-bleaching

Swann, T., Campbell, R., & Institute, T. A. (2016, June). *Great barrier bleached* (tech. rep.). Australian Institute. https://australiainstitute.org.au/wp-content/uploads/2020/12/Swann-Campbell-2016-Great-Barrier-Bleached-FINAL-w-cover.pdf

US Department of Commerce, N. O., & Administration, A. (n.d.). Noaa funded projects. Retrieved October 10, 2025, from https://coralreef.noaa.gov/conservation/funded_projects.html