THE ETHICS OF AUTOMATED CLASSIFICATION

A case study using a dignity lens

SmartyGrants
Software, data science & grantmaking intelligence

An enterprise of:
ourcommunity.com.au
Where not-for-profits go for help

The ethics of automated classification: a case study using a dignity lens

A SmartyGrants white paper, March 2022

Authors: Paola Oliva-Altamirano and Lorenn Ruster

Reviewers: Nathan Mifsud, Sarah Barker, Stefanie Ball and Kathy Richardson
Edited by Kerryn Burgess, graphic design by Amy Johannsohn

Published by Our Community Pty Ltd, Melbourne, Victoria, Australia

About the authors

Paola Oliva-Altamirano, Director of Data Science, SmartyGrants Innovation Lab

A research scientist trained in astrophysics, Paola designs algorithms to improve understanding of social sector data, with the goal of facilitating human-centred artificial intelligence (AI) solutions. She is actively involved in AI ethics discussions at the Innovation Lab and as a member of the Standards Australia AI Committee.

Paola graduated in physics in Honduras and later completed a PhD and post-doc in astrophysics at Swinburne University of Technology, Melbourne. In 2016 she co-founded Astrophysics in Central America and the Caribbean (**Alpha-Cen**) to support students in developing countries pursuing careers in science. She is currently the organisation's vice-president. Paola has been with the SmartyGrants Innovation Lab since 2018.

Lorenn Ruster, PhD candidate, Australian National University School of Cybernetics; and responsible tech collaborator, Centre for Public Impact

Lorenn is a social justice–driven professional interested in the intersection of technology, cross-sector collaboration, impact, systems change and dignity. For 10 years, Lorenn was a strategy consultant, most recently a director at PwC's Indigenous Consulting, working on projects spanning human rights, Indigenous co-design, collective impact and creating learning organisations. As an Acumen Fellow, she spent a year working with a Ugandan solar energy company as its marketing and innovation director, and as an alumnus of Singularity University's Global Solutions Program, she prototyped a device leveraging sensor technology for community-led landmine detection. In addition to her Masters in Applied Cybernetics from ANU, she holds a Masters in International Management (CEMS MIM) from the University of Sydney, Copenhagen Business School and HEC Paris. She also has a Bachelor of Science (Adv)/ Bachelor of Arts majoring in chemistry, psychology and French. Her doctoral studies at ANU's School of Cybernetics focus on the relationship between dignity and AI development.

Executive summary

SmartyGrants launched the text auto-classification system CLASSIEfier in 2021 to classify grantmaking records on behalf of grantmakers and other social sector supporters, with a view to tracking the flow of money in Australia by sector, location and beneficiary. The algorithm became a pilot for ethical considerations in artificial intelligence (AI) systems. At each stage of development, the team at the Innovation Lab considered the implications of their decisions and sought the best way to enable values that mattered, including (but not limited to) transparency, explainability, interpretability, equity and fairness.

As part of this work, the SmartyGrants Innovation Lab evaluated the dignity lens analytic tool (released by the Centre for Public Impact in 2021) as an ethics framework. The tool helped us audit each decision made in CLASSIEfier's development according to Donna Hicks' 10 essential elements of dignity. It also categorised each decision as protecting people from dignity violations or promoting dignity, or both. We found this distinction useful and suggest that the dignity lens analytic tool could work as a guideline in the development of AI and other data-driven products, and improve the documentation of AI-assisted decision-making. This white paper demonstrates how we used the dignity lens retrospectively; that is, after decisions had already been made. In the future, we expect to use it earlier in the AI development process, as a planning and design tool.

Introduction

CLASSIEfier is a keyword-matching model SmartyGrants uses to classify grants in Australia. One of its functions is to enable more informative tracking of funds across the Australian grantmaking landscape. The algorithm classifies grants using the social sector taxonomy **CLASSIE** and has expanded to include the **Sustainable Development Goals (SDGs)**.

When building the algorithm, we faced several ethical considerations. For example, what is the correct way to handle grant data without breaching confidentiality and data privacy? What degree of model accuracy is acceptable? How do we overcome human, data and algorithm bias? How involved should data experts and data owners be?

The ripple effect of classification systems in different industries varies according to their applications, but ultimately, these systems affect human users' decision-making. CLASSIEfier serves a significant proportion of grantmakers in Australia, auto-classifying more than 1 million grants in the SmartyGrants database. The data classification enables individual grantmakers to understand their funding distribution and its alignment to the social change they are trying to bring about. They can measure the fraction of funding allocated to specific subjects and populations, to evaluate impact, and to easily identify grants serving specific subjects, to list a few examples. In the short term, all this information can be used to prioritise applications for assessment, and in the long term, it can be used to assist in the planning of programs.

These potential benefits of CLASSIEfier have corresponding risks. Many risks stem from the possibility of incorrect classification, which leads to potentially misleading funding distribution outputs and incorrect impact evaluation data. Many risks stem from the possibility of incorrect classification or missing labels. This leads to potentially misleading funding distribution outputs and incorrect impact evaluation data. Given that the funding distribution information may be used by grantmakers to understand where best to allocate their money next, errors in this information could mean that good grant applicants miss out on funding. Further, incorrect classification of grants has flow-on effects on the validity of impact evaluations and decisions that use the aggregated data. The algorithm may influence decision-making on a large scale, with large numbers of grantmakers using the data, which means the potential impact is further magnified.

To improve the performance of the algorithm, the Innovation Lab has taken several steps to facilitate transparency, explainability, interpretability, stakeholder engagement, testing and incorporation of feedback. This white paper frames the decisions we made in terms of their impact on dignity, using the dignity lens analytic tool developed by Lorenn Ruster and Thea Snow and published in partnership with the Centre for Public Impact.

The dignity lens analytic tool

What is meant by dignity?

Ruster & Snow (2021) proposed this working definition of dignity:

> ***Dignity refers to the inherent value and inherent vulnerability of individuals. This worth is not connected to usefulness; it is equal amongst all humans from birth regardless of identity, ethnicity, religion, ability or any other factor. Dignity is a desire to be seen, heard, listened to and treated fairly; to be recognised, understood and to feel safe in the world. Dignity is influenced in positive and negative ways by others' behaviours and/or by technologies and other factors and at the same time, people have inviolable dignity.***

Ruster and Snow's definition adopts the 10 essential elements of dignity proposed by Donna Hicks (2013) to operationalise what dignity looks and feels like.

These elements are in the table below.

10 essential elements of dignity

1. **Acceptance of identity:** Having our identity accepted, no matter who we are
2. **Recognition:** recognition of our unique qualities and ways of life
3. **Acknowledgement:** being seen, heard, validated and responded to
4. **Inclusion:** having a sense of belonging, and feeling included at all levels of relationship (family, community, organisation and nation)
5. **Safety:** being physically and psychologically safe and secure
6. **Fairness:** being treated in a fair and even-handed way
7. **Independence:** feeling in control of life and experiencing a sense of hope and possibility
8. **Understanding:** actively listening, being given the chance to share perspectives
9. **Benefit of the doubt:** treating people as if they are trustworthy and operate with integrity
10. **Accountability:** taking responsibility for actions, apologising when harm has been done and committing to change hurtful behaviour.

What does a dignity ecosystem look like?

Ruster & Snow (2021) propose thinking about dignity as an ecosystem. This dynamic view of dignity captures the different roles that individuals and organisations can play in relation to dignity. These include:

- protective roles – mechanisms and actions that prevent dignity violations, or that remedy dignity violations if they do occur
- proactive roles – mechanisms and actions associated with promoting dignity.

All roles are underpinned by acknowledging dignity. Ruster & Snow (2021) believe that organisations need to play both protective and proactive roles to keep the dignity ecosystem in balance (see Figure 1).

Dignity Ecosystem

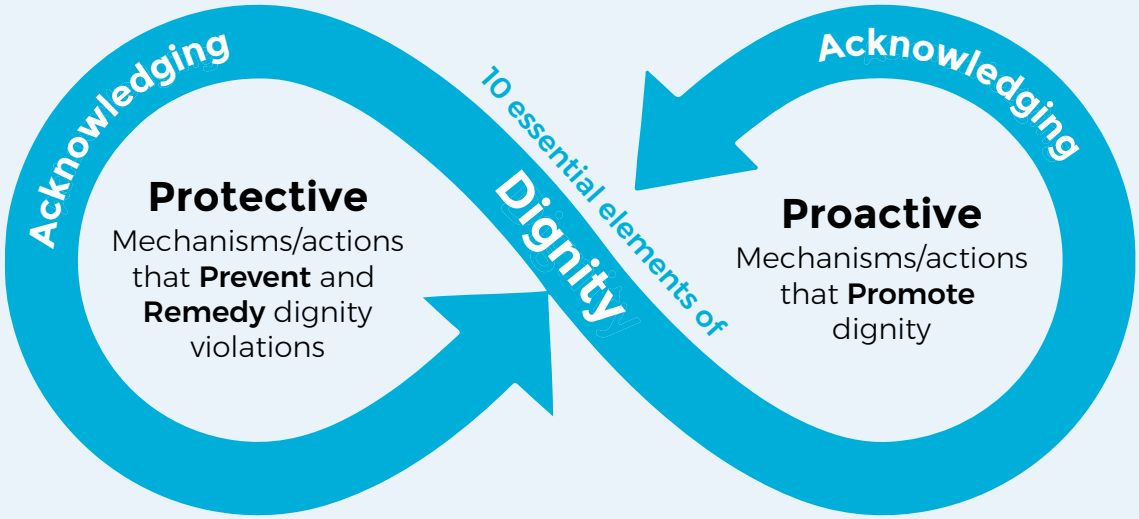


Figure 1: The dignity ecosystem (Ruster & Snow, 2020, page 9)

Making dignity real: applying the dignity lens

The dignity lens is an analysis tool to help organisations understand:

- which elements of dignity are reflected
- which types of roles we are playing (protective, proactive or both)
- and ultimately, what might we need to do to have a more balanced dignity ecosystem.

The Dignity Lens

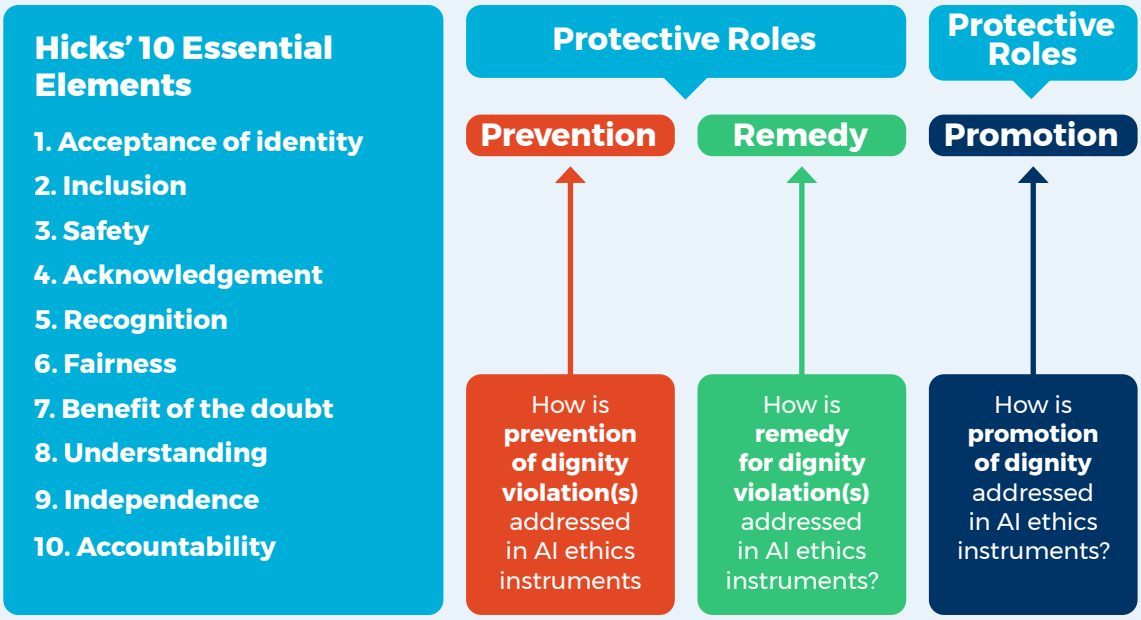


Figure 2: A schematic of the dignity lens applied (adapted from Ruster & Snow, 2021, page 10)

The dignity lens can be applied at various stages of development of the AI system. For example:

- **In planning and design of a new product, tool or initiative:** Apply the dignity lens to understand how the design upholds the essential elements of dignity and what could be done to strengthen dignity in the design. This is how the Innovation Lab plans to use the tool in the future.
- **In development (including data exploration, data modelling, deployment, and other processes):** Implement dignity review points to understand how balanced the dignity ecosystem is as you make the design into reality.
- **In testing:** Proactively scope what dignity means for the stakeholder groups involved in your testing processes and integrate these considerations into your testing and refining regime.
- **In release:** Consider the potential impacts of the implementation of the product, tool or initiative on the dignity of potential users, including how the product etc may integrate with other systems.
- **In review or monitoring:** Include the dignity lens as an assessment tool to understand where there may be risks to dignity, or whether there are opportunities to strengthen dignity in later releases. It can also be applied to decisions that have already been made, with a view to revisiting those decisions. This is how the Innovation Lab used the tool as presented in this whitepaper.
- **In de-commissioning:** Use the dignity tool to assess whether the product, tool or initiative needs to be shut down.

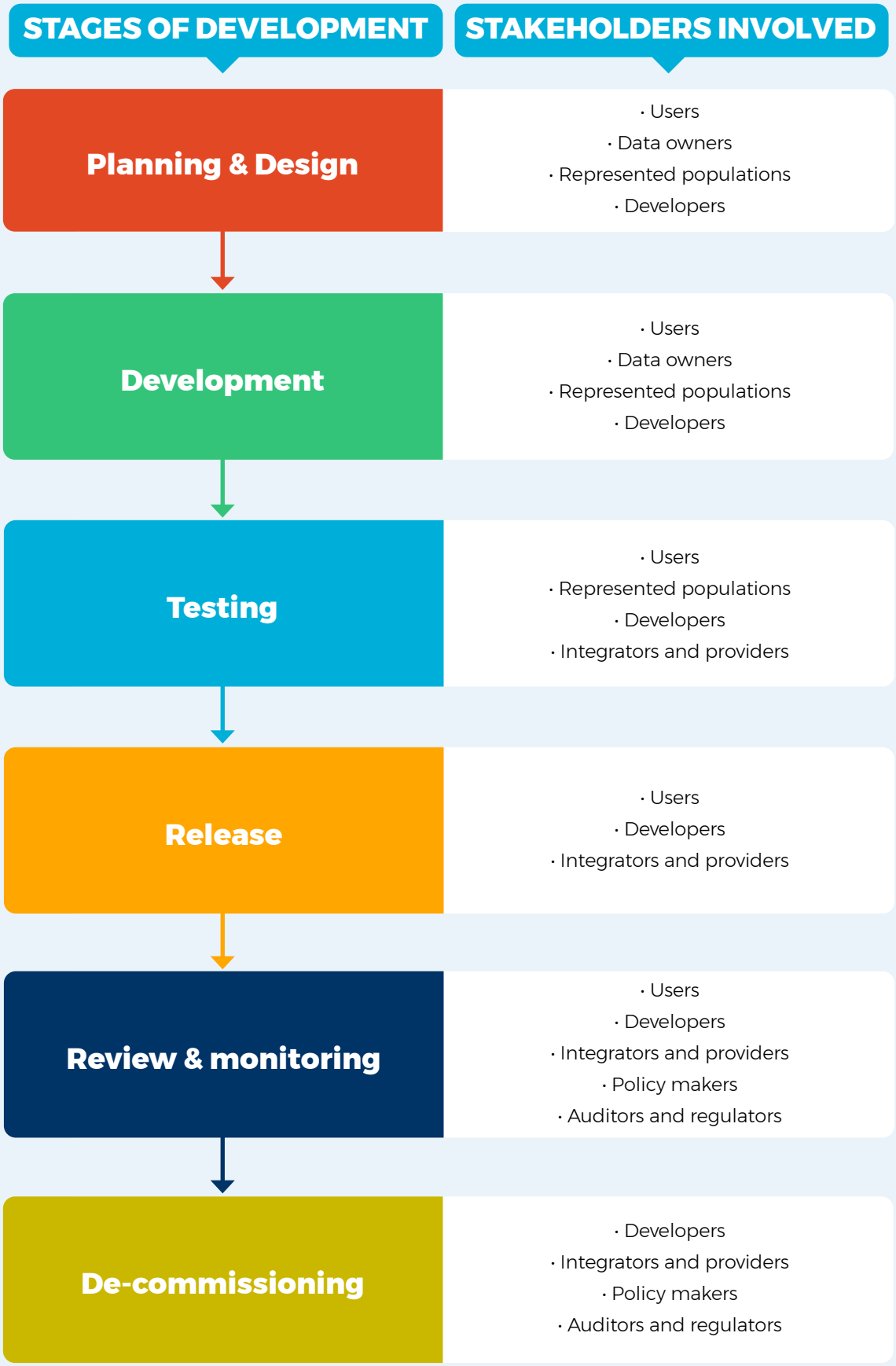
Different elements of dignity relate to different stakeholders involved in the development of AI systems in different ways. For example, when ensuring data privacy, we are protecting the safety of data owners. And when we invest in user experience (UX) design, we could be ensuring inclusion of diverse users.

The stakeholders who play a part in the various stages of development of an AI system include:

- **Users:** People who will use the AI system after creation.
- **Data owners:** People who have explicit rights over the use of the data. For example, in CLASSIEfier's case, the grant applicants own their grant applications and the grantmakers have the right to use the grant applications for reporting purposes. Both grant applicants and grantmakers are data owners.
- **Represented populations:** People represented in the data used to train the AI system.
- **Developers (including data scientists, data engineers, software developers, user experience designers, domain experts, and others):** The people involved in the design and creation of the AI system.
- **Integrators and providers:** People who will make the AI system available to the users.
- **Policy makers:** People who will use information provided by an AI system to guide their decision-making regarding the formulation of policies.
- **Auditors and regulators:** People who will evaluate the involvement of the AI system in decision making and the risks and benefits brought to society.

When developing AI systems, the designers, developers, integrators, providers, policy makers, auditors and regulators should work together to enable a balanced dignity ecosystem for users, data owners and the populations represented in the data.

Stakeholder involvement



3: Diagram to show the stakeholders involved in each stage of development when building AI systems

CLASSIEfier's dignity lens

In the following tables we explore the different stages of development of CLASSIEfier, the decisions made, and the elements of dignity upheld in those decisions. Each scenario refers to specific stakeholders.

Table 1: Consideration of the key activities and decisions made during the development of CLASSIEfier. We link each of the decisions made to Donna Hicks' elements of dignity as well as to the nature of the mechanism/action: protective (prevention and remedy) or proactive (promotion).

In this phase:	A decision was made to:	The elements of dignity upheld through this decision included:	The activities were protective because:	The activities were proactive because:
Planning and data exploration	Undertake scenario planning to anticipate what could go wrong (e.g. the possible effects of missing/wrong classifications, what happens when the data is not complete).	***Recognition*** We had a deep awareness that the algorithm depends on data owned by grant applicants and grantmakers. Without it, the tool would not exist. ***Inclusion and fairness*** We considered scenarios from diverse groups of users, data owners and populations represented in the data and tried to anticipate what being treated in a fair and even-handed way would mean in the context of the algorithm. ***Understanding*** We held brainstorming sessions with domain experts and data scientists to understand the uses of the algorithm and its implications for the populations represented in the data.	Scenario planning was done with the intent to prevent dignity violations that could emerge from the use of CLASSIEfier in various scenarios.	
Planning & data exploration	Investigate data distribution across possible classification categories.	***Acceptance of identity*** Our investigation revealed an uneven distribution of represented populations and insufficient labelling that could lead to discrimination. Knowing that an uneven distribution of represented populations translates to significant algorithm bias once it is deployed at scale, we found that we needed an improved training dataset.	Data distribution was investigated to prevent dignity violations that could have occurred if the training dataset hadn't been improved.	

In this phase:	A decision was made to:	The elements of dignity upheld through this decision included:	The activities were protective because:	The activities were proactive because:
Planning & data exploration	Investigate data distribution across possible classification categories.	***Fairness*** We investigated fairness across different represented populations by looking at data distribution across possible classification categories.		
	Train the model without using personal data.	***Safety*** We followed the **Innovation Lab data science guidelines** to prevent violation of stakeholder privacy and keep data secure.	Training the model without using personal data prevents stakeholder data privacy breaches, thereby protecting stakeholders from potential dignity violations.	
	Seek consent from data owners.	***Fairness and understanding*** We asked for consent from data owners (SmartyGrants clients) to use the data that would be needed for algorithmic testing.	Seeking consent prevents dignity violations by upholding ethical use of data, for example when it comes to privacy.	Seeking consent promotes dignity by giving data owners an opportunity to be heard and considered before a product is built.
Development	Prepare a training dataset which attempts to mitigate for identified data biases.	***Acceptance of identity*** We acknowledged the importance of fairly representing all subjects and populations when training the algorithm ***Inclusion and fairness*** We chose to mitigate bias by using only SmartyGrants data in the training dataset. Although public data appeared useful on face value, on closer examination we found that the bias inherent in the public data could harm the dignity of the populations represented in the data and the data owners. For example, data coming from news can unfairly link specific populations to alcohol consumption, family violence etc. ***Understanding*** We analysed the different outcomes generated by different training datasets, and their implications for stakeholders' dignity. We adjusted the model to account for our findings	Mitigating against identified biases in the training dataset **prevents** potential future dignity violations.	

In this phase:	A decision was made to:	The elements of dignity upheld through this decision included:	The activities were protective because:	The activities were proactive because:
Development	Abandon an algorithm which uses machine learning to classify data and instead create one that uses keyword-matching rules.	***Acceptance of identity*** We analysed the differences between machine learning and keyword-matching rules when classifying grants in terms of their impacts on different users and data owners. In doing so, we identified that machine learning was not the right solution to our problem because while it performed well for some subjects and populations it could also greatly harm others by overlooking or wrongly classifying unrepresented groups. ***Inclusion and Fairness*** Our decision to use keyword-matching algorithm allows transparency and higher accuracy when classifying grants. We also considered the different keywords from multiple cultural perspectives.	Choosing to go with keyword-matching rules instead of machine learning allowed for higher accuracy, transparency and explainability, preventing potential dignity violations. that could emerge from more 'black box' algorithms.	
	Ask internal and external reviewers to review the keywords used by the algorithm.	***Acceptance of identity and recognition*** We acted on our belief that the key-words used by the algorithm should accurately represent the subjects and populations in each CLASSIE category. ***Acknowledgement*** We acknowledged that the algorithm developers have limited contextual understanding. We sought external help to accurately represent different populations and data owners. ***Inclusion*** Consideration of what keywords could mean from different perspectives. ***Independence*** We acted on our belief that subject matter experts and data owners should have some control over what is produced and gave them the option of providing constructive feedback on the algorithm design, particularly the keywords used. ***Accountability*** CLASSIEfier (and its developers) was held accountable for the keywords used in the classification		Diverse feedback regarding the keywords used in design of the algorithm **was proactively collected** from users and developers, which will improve CLASSIEfier's performance and promote the dignity of subjects and populations represented in the data.

In this phase:	A decision was made to:	The elements of dignity upheld through this decision included:	The activities were protective because:	The activities were proactive because:
Testing	Promote CLASSIEfier in conferences, meetups, and other forums.	***Recognition*** CLASSIEfier was brought to the public to trigger discussion among data, domain, and design experts, and in this way their knowledge and expertise were recognised. ***Acknowledgement*** We acknowledged that the algorithm developers have limited domain expertise and external help was needed to mitigate algorithmic bias and build our capability. ***Understanding and accountability*** CLASSIEfier's developers were held accountable for the decisions made when building the algorithm through the testing process. Feedback was sought as a way of understanding unique perspectives and was incorporated when appropriate.		Diverse feedback was **proactively collected** from the public, data, domain and design experts, which will improve CLASSIEfier's performance and promote the dignity of subjects and populations represented in the data.
	Work closely with diverse data owners to test CLASSIEfier's performance.	***Acceptance of identity and recognition*** Data owners (testers) offered feedback regarding how their data was classified as part of the testing phase. ***Acknowledgement*** We acknowledged that the algorithm developers have limited domain expertise and that we needed to incorporate the data owners' perspective during delivery. ***Inclusion and fairness*** We considered testing scenarios from diverse groups of data owners and what fairness could look like for them. We also collected feedback on the possible uses of CLASSIEfier in the field and potential pitfalls. ***Independence*** We enabled the data owners to have control over the algorithm and their own data. ***Understanding and accountability*** We collected feedback from the data owners regarding their needs and then were held to account to implement the feedback gathered.	Testing CLASSIEfier against real use cases was done to identify any potential ways in which CLASSIEfier could cause harm to the data owners and **prevent** this from happening through improved algorithm performance and mitigating identified biases.	

In this phase:	A decision was made to:	The elements of dignity upheld through this decision included:	The activities were protective because:	The activities were proactive because:
Testing	Open the algorithm's code for review, and invite data scientist Kabir Manandhar Shrestha to join the Innovation Lab for three months to review CLASSIEfier. His review is summarised in the article "**Ethical considerations in multilabel text classifications**".	***Acknowledgement and fairness*** We acknowledged that the algorithm developers had limited expertise and welcomed an external reviewer to identify and mitigate bias. ***Understanding and accountability*** We collected feedback from the external reviewer and then were held to account to implement the feedback gathered.	Reviewing CLASSIEfier from an external data scientist's perspective was helpful in identifying bias and potential mitigations, **preventing** potential violations of the dignity of users if the bias had been left unaddressed.	
Release	Integrate CLASSIEfier into SmartyGrants, enabling users to customise the tool according to their needs.	***Acceptance of identity, recognition, inclusion*** When integrating CLASSIEfier into SmartyGrants we considered diverse scenarios from different users' perspective. In doing so, we enabled users to customise the tool according to their needs. ***Independence*** Through the integration with SmartyGrants, users can choose which text should be classified, the level of classification and the maximum number of labels accepted. This allows for user control and helps avoid noisy data and unwanted labels.		Allowing users to customise the tool according to their needs promotes the dignity of users.
	Publish the results in plain English.	***Inclusion and accountability*** We tried to cater to diverse audiences by publishing the results using plain and simple English and using clear and understandable visualisations. In doing this, we aimed to increase the likelihood that people would use the outputs to inform their own work. Thus we promoted not only inclusion but also greater accountability – if more people can understand the results, more can hold us to account.		Publishing the results in plain English promotes dignity by maximising accessibility. It enables people to hold us to account and to use the outputs to inform their own work.

In this phase:	A decision was made to:	The elements of dignity upheld through this decision included:	The activities were protective because:	The activities were proactive because:
Release	Publish the results in aggregate form.	***Safety*** The results were presented in aggregate form to avoid harms that could emerge from data identification.	Publishing aggregate results **prevents** dignity violations that could emerge from data identification.	
Review and monitoring	Make CLASSIEfier a live tool, open to feedback.	***Acknowledgement, understanding and accountability*** We are actively collecting and incorporating feedback from users on an ongoing basis. For example, data users and owners can suggest changes to the keywords, the system integration and the user interface(s). This feedback mechanism not only acknowledges their different experiences but also gives them a way of taking control of their experience. In this way we continue to seek understanding of different users' experiences of the tool. Improvements are released iteratively based on lessons learnt.		Mechanisms for continual feedback enable dignity promotion – users are seen, heard and listened to on a regular basis.

Table 2: Mechanisms we enacted, mapped to Donna Hicks' 10 essential elements of dignity, a protective/proactive stance and the stages of AI development

Donna Hicks 10 Essential Elements of Dignity	Protective /Proactive	Stages of AI development				
		Planning & data exploration	Development	Testing	Release	Review & monitoring
1. Acceptance of identity Approach people as being neither inferior nor superior to you; give others the freedom to express their authentic selves without fear of being negatively judged; interact without prejudice or bias, accepting that characteristics such as race, religion, gender, class, sexual orientation, age, and disability are at the core of their identities.	**Protective**	Data distribution investigation	Preparing a training data-set to mitigate against identi-fied biases Model assessment and selection aligned to needs	Testing with data owners		
	Proactive		Feedback on keywords used in the model		Customised user interface	
2. Recognition Validate others for their talents, hard work, thoughtfulness, and help; be generous with praise; give credit to others for their contributions, ideas, and experiences.	**Protective**	Scenario planning		Testing with data owners		
	Proactive		Feedback on keywords used in the model	Feedback on CLASSIEfier in diverse forums	Customised user interface	
3. Acknowledgment Give people your full attention by listening, hearing, validating, and responding to their concerns and what they have been through.	**Protective**			Testing with data owners Algorithmic review by external party		
	Proactive		Feedback on keywords used in the model	Feedback on CLASSIEfier in diverse forums		Mechanisms for continual feedback

Donna Hicks 10 Essential Elements of Dignity	Protective /Proactive	Stages of AI development				
		Planning & data exploration	Development	Testing	Release	Review & monitoring
4. Inclusion Make others feel that they belong, at all levels of relationship (family, community, organization, and nation).	**Protective**	Scenario planning	Mitigation against identified biases Adoption of keyword-matching rules instead of machine learning	Testing with data owners		
	Proactive		Feedback on keywords used in the model		Customised user interface Results published in plain English	
5. Safety Put people at ease at two levels: physically, so they feel free from the possibility of bodily harm, and psychologically, so they feel free from concern about being shamed or humiliated and free to speak without fear of retribution.	**Protective**	Model trained without personal data			Results published as aggregate data	
	Proactive					
6. Fairness Treat people justly, with equality, and in an even-handed way according to agreed-on laws and rules.	**Protective**	Scenario planning Data distribution investigation Consent for data use	Mitigation against identified biases Adoption of keyword-matching rules instead of machine learning	Testing with data owners Algorithmic review by external party		
	Proactive	Consent for data use				

Donna Hicks 10 Essential Elements of Dignity	Protective /Proactive	Stages of AI development				
		Planning & data exploration	Development	Testing	Release	Review & monitoring
7. Independence Encourage people to act on their own behalf so that they feel in control of their lives and experience a sense of hope and possibility.	**Protective**			Testing with data owners		
	Proactive		Feedback on keywords used in model		Customised user interface	
8. Understanding Believe that what others think matters; give them the chance to explain their perspectives and express their points of view; actively listen in order to understand them.	**Protective**	Scenario planning Consent for data use	Mitigation against identified biases	Testing with data owners Algorithmic review by external party		
	Proactive	Consent for data use		Feedback on CLASSIEfier in diverse forums		Mechanisms for continual feedback
9. Benefit of the doubt Treat people as if they are trustworthy; start with the premise that others have good motives and are acting with integrity.	**Protective**					
	Proactive					
10. Accountability Take responsibility for your actions; apologize if you have violated another person's dignity; make a commitment to change hurtful behaviours.	**Protective**			Testing with data owners Algorithmic review by external party		
	Proactive		Feedback on keywords used in the model	Feedback on CLASSIEfier in diverse forums	Results published in plain English	Mechanisms for continual feedback

Table 3: Overview of protective and proactive mechanisms identified

Protective mechanisms	Proactive mechanisms
Bias-related mechanisms · Data distribution investigation · Mitigation against identified biases **Model assessment and selection aligned to needs** **Testing with data owners** **Scenario planning** **Anonymisation** · Model trained without personal data · Results published as aggregate data **Algorithmic review by external party**	**Feedback mechanisms:** · Feedback on keywords used in the model · Feedback on CLASSIEfier in diverse forums · Mechanisms for continual feedback once 'live' **Customised user interface** **Consent for data use** **Results published in plain English**

Reflections on the use of the tool

During the development of CLASSIEfier, the Innovation Lab made decisions through brainstorming and team consensus. Our collaborative approach was quite natural to the team, but we were curious about what we might be missing and how we could be more rigorous in considering dignity. We adopted the dignity lens analytic tool after CLASSIEfier was ready for public release, as a way of reviewing what we had done and to ensure we documented our decisions. The framework has been a valuable tool for documenting the ethical questions we faced and the resolutions we took. We found that the tool can adapt to auto-classification and artificial intelligence systems as well as other data-driven projects, such as data visualisation, insight reports, survey design and more.

Although the dignity lens analytic tool can be used retrospectively, as was done with CLASSIEfier, we believe that using the tool in a proactive, prospective way, upfront in the design process, would be even more helpful. A few ways in which we think it could be helpful are outlined below.

The dignity lens assists us to achieve a balance between protective and proactive mechanisms

Ensuring a balance between protective and proactive mechanisms is a core part of the dignity lens. We found more protective mechanisms than proactive ones (see Table 3). By using this tool upfront, we believe we would have been able to identify more ways of proactively incorporating dignity into our design process and our end product. For example, as it stands, we have not identified any proactive mechanisms for the element of safety. Considering what this means for our different stakeholder groups upfront could yield some interesting mechanism additions. It should also be noted that mechanisms can be applied in both protective and proactive ways. For example, scenario planning for us was all about preventing potential harms to dignity, but we could also have used scenario planning in a proactive way, to think about how we might enable dignity in the different scenarios identified.

The dignity lens helps us address all 10 essential elements of dignity

We addressed some aspects of dignity more comprehensively than others. For example, given that we were already deeply considering how to identify and mitigate bias, elements such as “acceptance of identity”, “inclusion”, “fairness” and “understanding” were addressed as part of the bias mitigation process. Similarly, a range of feedback and testing mechanisms allowed us to address elements such as “recognition”, “acknowledgement”, “understanding” and “accountability”. However, the analysis shows that some elements of dignity were less obvious to us, such as the element “benefit of the doubt”, against which we have no mechanisms mapped. Thinking about this element upfront could have yielded some interesting design decisions. For example, perhaps consideration of “benefit of the doubt” could have led to encouraged debate about the governance surrounding the use of the model in ‘edge cases’ (situations that would occur when using the AI system at extreme parameters).

The dignity lens enables us to give adequate consideration to all stages of AI development

The initial version of the dignity lens did not directly consider the different stages of AI development; we adapted it through the course of implementing the tool. The analysis shows that, overall, we had more mechanisms operating in the earlier stages of AI development than at later stages (see Table 2). Active consideration of this in the design phase could yield more mechanisms that address the release stage and the review and monitoring stage and lead to discussions about de-commissioning, which were not considered upfront by our team. Even though our team valued feedback, we did not employ formal feedback mechanisms in the planning phase. Thinking about feedback upfront would assist us in recognising that the expertise of stakeholders such as data experts and data owners needs to be incorporated before we even develop a model. Similarly, we might consider protective accountability mechanisms at other stages of the AI development process, for example during development (so that changes can be implemented before testing).

The dignity lens provides a language for discussion and debate

We reflected that using the dignity lens provided a language for discussing the challenges and tensions that we faced in the design and implementation process. We believe that having this language in use earlier would assist with conversations, not only amongst the Innovation Lab team but also with other stakeholders, including potential users of the product and data owners.

The dignity lens gives us confidence and a way of documenting our decisions so we can continually improve

We pride ourselves on being ethical in our approach, but of course it is difficult to verify that we are 'walking the talk'. Using the dignity lens helped us solidify our own decisions and processes. We believe using the dignity lens upfront would give us even more confidence that we are actively and intentionally thinking about and incorporating dignity considerations throughout the stages of AI development. Before using it, we believed we were upholding dignity, but we had no documented process to show it. Now we have a way of systematically reflecting on and documenting our processes, enabling us to improve for next time.

In future, the Innovation Lab intends to use a proactive approach to considering dignity in producing data-driven tools. The dignity lens will be used during planning and building to guide ethical considerations of scoping and design of our products.

Contributors

SmartyGrants

SmartyGrants is a grants management system that allows grantmakers to receive and manage applications – but it is more than just a tech solution. SmartyGrants drives sector-wide reform by building best practice into an intuitive and affordable product that grantmakers want to use because it makes their lives easier and their outcomes better.

We are extending our product with added intelligence and insights. We want to help grantmakers become more *efficient* and *effective* by enabling decision making that is driven by data and outcomes. These improvements will benefit the community by ensuring money is going where it's needed and the best projects get funded.

Centre for Public Impact

The Centre for Public Impact is a global not-for-profit founded by the Boston Consulting group. It acts as a learning partner for governments, public servants, and the diverse network of changemakers leading the charge to reimagine government so that it works for everyone.

School of Cybernetics, Australian National University

The School of Cybernetics is part of the College of Engineering and Computer Science at the Australian National University. It is focused on guiding and accelerating a new branch of engineering centred on the safe, sustainable and responsible scale of cyber-physical systems and artificial intelligence.

Bibliography

Hicks, D, 2013, *Dignity: the essential role it plays in resolving conflict in our lives and relationships*, Yale University Press, New Haven, Conn.; London.

Manandhar Shrestha K, 2021, *Ethical considerations in multilabel text classifications. An Australian community sector case study on the quest for explainable AI.* SmartyGrants.com.au **https://smartygrants.com.au/uploads/general/SG/Research/WhitePaper_Ethical-considerations-in-multilabel-text-classifications.pdf**

Ruster, LP & Snow, T, 2021, *Exploring the role of dignity in government AI Ethics instruments*, Centre for Public impact, Viewed 19 March 2021, **https://www.centreforpublicimpact.org/partnering-for-learning/cultivating-a-dignity-ecosystem-in-government-ai-ethics-instruments**.

NOTES