A Interactive Strategy List with Explanations

A full interactive version of the strategy list including explanations per strategy can be found on https://explainable-ai.nl/4d/.

B Survey to define scope and relevant values

B.1 Method an Participants

We surveyed 36 subjects at a Dutch insurance association's 'Data Afternoon 2024' to assess the relevance of the values 'fairness' and 'wealth' for digital price differentiation and to explore which other values are deemed relevant. The survey had three questions:

(1) The value 'fairness' is important for justifying digital differentiation? (1 = strongly disagree, 3 = neutral, 5 = strongly agree)
(2) The value 'wealth' is important for justifying digital differentiation? (1 = strongly disagree, 3 = neutral, 5 = strongly agree)
(3) Which values are important to you to justify digital differentiation?

The survey was conducted in Dutch and participants answered the questions on their phone. Question (1) and (2) where shown at the same time. The survey was anonymous but subjects entered their job title or role. Table 1 provides an overview of the number of participants for each role. The 4 participants classifed as'other' described themselves as salesperson, manager, business analyst. and general insurance practisionar; the 7 'undefined' roles had entered their name instead of a role or an unclear abbreviation.

B.2 Results and Interpretation for Selecting Relevant Literature

Our survey results indicate that wealth and fairness are considered significantly more important than a neutral stance (score of 3) for justifying digital price differentiation, as reflected in responses to questions (1) and (2). Figure 2 illustrates how strongly participants agree with the importance of these factors. Specifically,

- The 36 particpants had a mean importance of wealth for justifying digital price differentiation of 3.65 ($SD = 0.92$), which is significantly above the neutral score of 3, $t(35) = 4.1133$, $p = 0.0002$.
- The 36 particpants had a mean importance of fairness for justifying digital price differentiation of 4.1 ($SD = 0.66$), which is also significantly above the neutral score of 3, $t(35) = 10.1730$, $p = 0.0001$

These results support our decision to focus this study on wealth and fairness as key factors in justifying digital price diffentiation. Consequently, we prioritize these two values in our literature selection to align with practical needs.

Furthermore, fairness is perceived as slightly, yet significantly, more important than wealth for justifying digital price diffentiation. The 34 participants had an average difference between fairness and wealth of 0.5 ($SD = 0.96$) indicating a small but very significant higher importance of fairness, $t(33) = 3.033$, $p = 0.0005$. This aligns with prevailing industry perspectives, where wealth optimization is often assumed, while fairness remains a subject of greater scrutiny. In other words, the emphasis on fairness in contemporary discussions may prime participants to prioritize it. Accordingly, Section 3.2 focuses the literature selection on fairness, assuming that wealth optimization

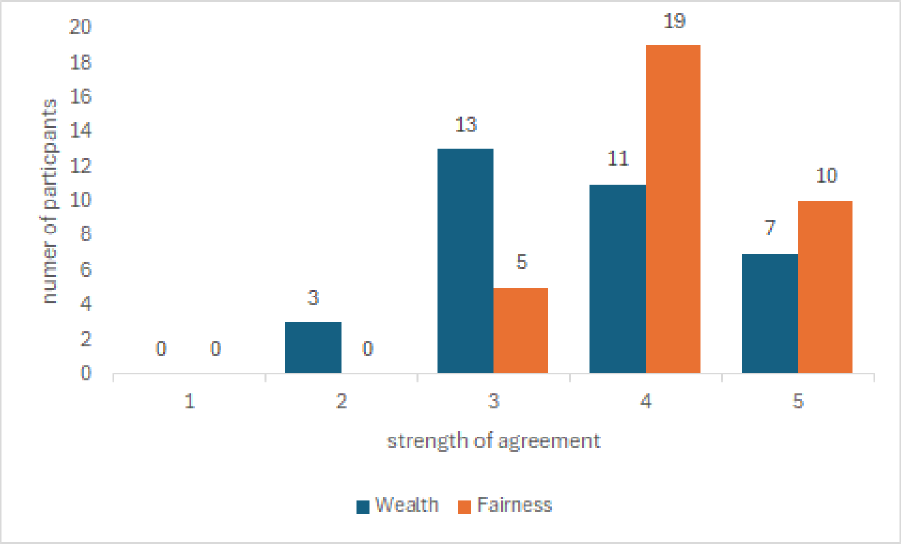

Fig. 2. Bar chart depicting the perceived importance of fairness (orange) and wealth (blue) in justifying digital price differentiation. Scores range from 1 = strongly disagree to 5 = strongly agree, with 3 representing a neutral stance.

is inherently addressed through price differentiation. Section 5 further examines how our list of strategies and selected literature reflect this emphasis.

Beyond fairness and wealth, our participants identified 17 distinct values relevant to justifying digital price differentiation in response to question (3) (see Table 2). Solidarity, transparency, and explainability were mentioned most frequently. The prominence of solidarity aligns with its recognized importance in the Dutch Association of Insurers' Code of Conduct 2018 [30] and the Solidarity Monitor [48]. Notably, Insurers [30] describe solidarity as instrumental to social responsibility and not a value on its own. Moreover, in legislative discourse, the term solidarity is increasingly replaced by fairness (see Section 2). Interviews with professionals involved in digital price differentiation (see Section 3.5) confirmed that solidarity is often perceived as a vague or outdated term compared to fairness. We therefore conclude that while solidarity remains relevant for literature selection, and should be included in our query, fairness should be the primary focus of this study.

Other values identified, such as transparency and explainability, were not included in our core literature query but remain instrumental in justifying fairness. Transparent and explainable processes facilitate clearer communication and justification of digital price differentiation decisions. Their importance is reflected in several strategies (e.g., Strategy 25 in Table 4) and further explored in Section 5. The remaining values were mentioned infrequently (fewer than four times) and are either encompassed by fairness (e.g., just) and wealth (e.g., accuracy, market demand), or they do not refer to actual abstract values, as defined by Oosterlaken [37] (e.g., behavior).

Function	N
Actuary	6
Consultant	3
Compliance Officer	4
Data Scientist	9
IT Engineer	3
Other	4
Undefined	7
Total	**36**

Table 1. Distribution of functions of the participants of the survey.

Word Entered	Count
Solidarity	12
Transparency	7
Explainability	6
Behaviour	3
Fairness	3
Clarity	2
Just	1
Representative	1
Reliability	1
Support	1
Insensitive variables	1
Statistical Significance	1
Insurability	1
Openness	1
Accuracy	1
Granularity	1
Market Demand	1

Table 2. Overview of values mentioned as relevant to digital price differentiation by professionals in insurance.

C List of Strategies Organized into Five Phases

Category	Index	Strategy employed in the Understanding phase	Source
Discrimination & Bias	1	Understand the difference between direct discrimination and indirect discrimination and the related concepts of ‘proxy variable’ and ‘disparate impact’.	[7, 25]
	2	Understand the difference between two forms of bias: bias because the data does not accurately represent the world (statistical bias) or bias because the data does not represent the world as it ideally could or should be (societal bias).	[7]
Technology & Fairness	3	Understand that removing bias also removes information; information that (potentially) leads to an accurate risk assessment and at the same time (potentially) leads to an unfair premium distribution.	[36]
	4	Understand that there is no one size fits all for fairness and that technology, ethics or law will not provide a definitive answer as to what is best.	[7, 49]
	5	Understand that there is not a single threshold value for what is an acceptable amount of bias, or acceptable amount of disparate impact	[6]
	6	Understand that information about sensitive variables is necessary to measure fairness and is privacy-sensitive.	[14, 47]
Legal & Ethical	7	Understand that the AI Act, EU law and national law specify protected variables that may not be directly discriminated against (e.g., the grounds for discrimination in the AWGB), but do not provide clarity on which variables can be used without indirect discrimination.	[14, 51]
	8	Understand that the justification for making distinctions on a variable depends on the legal and ethical criteria in Table 8.	[3, 6, 21, 25, 32, 43, 50]
Insurance	9	Understand the importance of information asymmetry, especially how unequal knowledge about customers’ risk profiles can lead to adverse selection.	[10]
	10	Understand the difference between a risk-based premium and premiums based on non-risk factors (such as giving discounts to attract customers).	[25]

Table 3. Strategies for the Understanding phase.

Category	Index	Strategy employed in the Determination Phase	Source
Target audience	11	Determine the target group to whom you wish to offer the insurance and estimate the consequences for the return and fairness of this choice.	[25]
Determine high-level strategy	12	Determine the level of rigor in dealing with sensitive variables, with the two extremes being (1) not using directly only legally protected variables and (2) completely equalizing premiums across all customers.	[25]
	13	Determine whether the premium will be partly based on non-risk factors (such as giving discounts to attract customers).	[29, 51]
	14	Determine lower bounds for the expected fairness (given a fairness measure) and the expected return of the price differentiation.	[11]
Determine sensitive variables	15	Determine the sensitivity of a variable and the extent to which it is justified to use it given the ethical and legal criteria in Table 8.	[3, 6, 21, 25, 32, 43, 50]
Fairness measurement	16	Determine whether the risk labels in the training dataset are sufficiently trustworthy to base a fairness score on.	[2, 40]
	17	Determine whether the fairness measure measures ’individual fairness’ (equality between two matching individuals) or ’group fairness’ (equality between two matching groups).	[36]
	18	Determine what the fairness measure compares between two groups or individuals. Fairness measures differ in particular in whether they (1) compare the risk assessment itself, (2) the margin of error in risk assessment (focusing on overestimation or underestimation), and (3) the extent to which they control for whether differences are caused by other nonsensitive variables.	[2, 40]
	19	Determine whether the fairness measure uses protected variables directly to measure fairness or whether the measure uses an estimate of a protected variable based on proxies.	[47, 51]
	20	Determine the implications of your fairness measure choices for selecting a strategy to increase fairness in pre-processing, in-training or post-processing (see strategy 21 to 27).	[3, 6, 21, 25, 32, 43, 50]

Table 4. Strategies for the Determination phase.

Category	Index	Strategy employed in the Adjustment phase	Source
Pre-processing	21	Adjust dataset by removing sensitive variables and any variables or interactions of variables that correlate.	[6, 36, 51]
	22	Adjust dataset by correcting any correlation around a sensitive variable.	[6, 36, 51]
	23	Adjust dataset by adding synthetic data points that level out the sensitive relationships.	[6]
In-training	24	Adjust algorithm so that it optimizes on an accuracy score while keeping the fairness score above a certain threshold.	[6, 36]
	25	Adjusting the algorithm to make it explainable.	[6, 51]
Post-processing	26	Transforming the predicted risk assessment to increase fairness.	[6, 36, 51]

Table 5. Strategies for the Adjustments phase.

Category	Index	Strategy employed in the Evaluation phase	Source
Evaluation	27	Evaluate quantatively whether the adjusted prediction model exceeds the fairness and return threshold..	[51]
	28	Evalualate qualitatively whether the adjusted prediction model is fair given the ethical and legal criteria in Table 8.	[43]
	29	Evaluate the model repeatedly in practice (after launch) for fairness through audits, statistics, experiments or observational studies.	[6]

Table 6. Strategies for the Evaluation phase.

Category	Index	Strategy employed in the Communication phase	Source
Communication	30	Communicating to the customer the social role of an insurance company in supporting risk spreading.	[25]
	31	Communicating to the client the need for risk assessment both to be profitable and to give an ’appropriate’ premium for a risk.	[25]
	32	Communicate to regulators the choices regarding fairness: the estimation of sensitive variables, the fairness measure, the choice of a specific threshold value and the final model.	[3, 6, 21, 25, 32, 43, 50]
	33	Enabling customers, competitors and regulators to challenge the model by setting up procedures for doing so.	[7]

Table 7. Strategies for the Communication phase.

Category	Index	Principle	Source
Legal	1	The legality of a distinction depends on a *legitimate interest* for the distinction.	[18, 25, 35, 50]
	2	The legality of a distinction depends on whether the distinction is *necessary* to achieve the goal: the necessity of using the variable is to be assessed by analyzing whether there is another variable that achieves a similar result but has a less negative impact on fairness.	[21, 35, 50]
	3	The legality of a distinction depends on whether the insurer's goal is *proportionate* to the affected interests of the customers.	[35, 50]
Ethics	4	The *mutability* of a variable by the client partly determines how justified it is to make distinctions based on that variable.	[3, 25]
	5	The *statistic correlations* around a variable partly determine how justified it is to make distinctions based on this variable; in particular, the correlation with the predicted risk of a customer and the correlations with legally protected variables.	[25, 32]
	6	The *causal relationships* around a variable partly determine how justified it is to make distinctions based on this variable.	[25]
	7	The *prejudices from societal history* (such as around skin color) partly determine how justified it is to make distinctions based on this variable.	[25, 32]
	8	The *effect on desirable behaviour* partly determines how justified it is to make distinctions based on this variable.	[3, 25]

Table 8. Legal and ethical principles influencing the strategies.