Laboratoire d'Analyse et Modélisation
de Systèmes pour l'Aide à la Décision
UMR 7243

CAHIER DU LAMSADE

413
February, 2026

Explainability and justification of automatic-decision making
A conceptual framework and a practical application

Sarra Tajouri, Yves Meinard
Alexis Tsoukiàs, Thierry Kirat

Explainability and justification of automatic-decision making – A conceptual framework and a practical application

Sarra Tajouri[1], Yves Meinard[2], Alexis Tsoukiàs[1], and Thierry Kirat[3]

[1]LAMSADE-CNRS, Université Paris Dauphine - PSL
[2]Aix Marseille Univ, CNRS, CGGG, Centre Gilles Gaston Granger, Aix-en-Provence, France
[3]TRIANGLE-CNRS, ENS Lyon, Université Lumière-Lyon 2, Sciences Po Lyon, Université Jean Monnet-Saint-Etienne

February, 2026

Abstract

Explainability of algorithmic decision-making systems is both a regulatory objective and an area of intense research. The article argues that a crucial condition for the acceptability of algorithmic decision-making systems is that decisions must be justified in the eyes of their recipients. We make a clear distinction between explanation and justification. Explanations describe how a decision was made, while justifications give reasons that aim to make the decision acceptable. We propose a conceptual framework of explanations and justifications, based on Habermas's theory of communicative action and Perelman's New Rhetoric theory of law. This framework helps to analyze how different forms of explanation can support or fail to support justification. We illustrate our approach with a case study on university admissions in France.

1 Introduction

The use of decision-support systems has become increasingly prevalent in contemporary societies. Many people are subject to decisions made by these systems on a regular basis, often without even realizing it [16, 75]. Their influence spans a broad spectrum of contexts, from seemingly trivial everyday interactions, such as personalized recommendations for movies, food, or consumer products, to high-stakes decisions involving access to social benefits, creditworthiness assessment, healthcare diagnosis, and even predictive policing [19, 35, 69, 71].

In some particularly high-stakes domains, such as access to public or private resources, these decision-support systems can have profound impacts, including unintended consequences on people's lives and even communities [51]. For instance, credit scoring algorithms used by banks to assess loan applications can (presumably inadvertently) discriminate against applicants from minorities or with low-income backgrounds, denying them access to credit and perpetuating cycles of poverty [10, 41]. Another telling example is the one of predictive policing algorithms that identify "high-crime" areas and thereby can lead to over-policing of marginalized neighborhoods, exacerbating systemic racism [6]. Similarly, in healthcare, algorithms can assist in diagnosing diseases or prioritizing patients for treatment, directly affecting patient care outcomes [2, 61, 60]. Recommendation systems on social media platforms also shape public discourse by influencing the information that people can access [14, 47].

In most cases, such systems or tools are called "decision tools" through misuse of language, since what they really do is produce a recommendation that an agent or agents can use or ignore when making their decision. However, the fact that most of these systems do not make decisions themselves does not make those decisions any less consequential.

Algorithmic systems hence actively shape society. They participate in social and economic structures, entrench norms (e.g., norms of creditworthiness or risk assessment; [23, 43]), and even mediate interactions [25], thereby displaying attributes that typically rather characterize social actors, such as agency and influence on social relationships.

Consequently, it is now largely accepted that these systems must align with certain norms and expectations, typically expressed using terms like "fairness" or "explainability", and they are increasingly subject to accountability requirements similar to those expected from human social actors [4]. However, algorithms do not function in the same way as human decision-makers and, crucially, people perceive and judge algorithmic and human decisions in distinct ways [28, 58]. Among key-differences, a stark contrast, which is unusual in human decision contexts, often characterizes automated decisions. While the accuracy and efficiency of human decisions often go hand in hand with decision-makers' ability to account for the way decisions were

made, automated systems often manage to perform highly accurately and efficiently some tasks without anyone being able to fully understand the details of how these tasks were performed. This difficulty is compounded when such systems produce recommendations that have multifarious, diffuse impacts, such as reinforcing existing societal inequalities [12], or when they reflect the biases and assumptions of the model developers [56]. The challenge of adapting to automated decisions requirements originally applicable to human decisions has prompted legislators to elaborate legal frameworks and regulations intended to guide the use of at least partly automated decision systems [21, 20].

A related challenge echoes the above clarification that such systems are typically not decision makers, but recommendation systems. Because what they produce is recommendations, the usability of such systems and, in fine, their effectiveness, depend on recommendations being followed by decision-makers and their consequences accepted by people they affect. The challenge, which is also typically addressed in discussions articulated in the terms just mentioned ("explanability", "fairness", etc.) is hence to foster the appropriation and acceptability of such systems.

In this article, we elaborate a framework designed to streamline and rationalize such attempts at providing a normative structure to the use of, at least partly, automated decision support systems. Most of the attempts in that direction in the existing literature revolve around the terms "explanation" and "justification" (or associated terms such as "explanability" and the like), but these terms are given different meanings in different contexts, resulting in a blurry and unclear picture. To address this predicament, we propose a clarification of the corresponding notions.

The remainder of this paper is organized as follows. Section 2 reviews the academic literature to map the meanings given to the terms "explanation", "justification" and the like, particularly in their application to, at least partly, automated decision processes. As a preliminary step in this clarification, we briefly examine how European legislation addresses these concepts, particularly the prominence of explanation and the relative absence of any formal justification requirement. This legal backdrop helps highlight both the normative expectations currently placed on algorithmic systems and the conceptual gaps that remain unaddressed in regulation, thus reinforcing the need for a more precise theoretical account. We then propose in section 3 our own conceptual framework to clarify the concepts of explanation and justification, based on the work of contemporary philosopher Jurgen Habermas [33] and legal theorist Chaïm Perelman [63]. The last section 4 proposes an application of this conceptual framework in a concrete, real-life case-study.

2 “Explanation” and “justification” in the existing legislation and literature

As mentioned above, this section examines the main features of European legislation on explainability and highlights a gap with regard to the issue of justification. It then goes on to discuss in greater detail the existing types of explanation and their compatibility with justification.

2.1 Explainability requirements and justification gaps in European legislation

A prominent legal framework in the domain is the GDPR (General Data Protection Regulation) [21], which was enacted by the European Union in 2018. The GDPR includes a “right to explanation”, but fails to explicitly define what it calls an “explanation”. It establishes “transparency" requirements for automated decision-making. Articles 13 and 14 mandate that individuals should receive “meaningful information about the logic, significance, and consequences of decisions" based on automated systems. Article 22 grants individuals the “right not to be subject to fully automated decisions", while Recital 71 emphasizes the need for “fair" and “transparent" processing, suggesting a right to understand and challenge algorithmic decisions. However, some researchers [18, 73] question the feasibility of such a right, criticizing its lack of precise language. Additionally, it is important to note that this regulation applies only to fully automated decision-making systems, which are relatively rare in practice. Indeed a distinction must be made between a “decision" and a “recommendation". The former refers to an allocation of resources (of any type) to one or more tasks, formally defined as a partition of a set (of “possible allocations”) which actually changes the state of the system upon which the allocation acts [15]. By contrast, the latter is a mere suggestion that a decision (in the above sens) should be made.

More recently, the AI Act [20], proposed by the European Commission in 2021 and enacted in June, 2024, targets high-risk AI systems, demanding transparency and clear documentation of their decision-making processes. This regulation aims to ensure that both developers and users of AI systems offer comprehensible “explanations”, enhancing accountability and safeguarding individual rights in critical sectors such as healthcare, finance, and public administration. It mandates that providers implement a risk management system understood as a *continuous and iterative process to identify, analyze risks* that may arise from both intended use and unintended consequences (Article 9).

In order to maintain human control, the AI Act enforces human oversight mechanisms (Article 14) and makes mandatory for AI providers to both continuously monitor system performance and report incidents (Article 61). Finally, strict enforcement

measures, including financial fines (Articles 85-92), ensure that AI developers remain legally accountable for the societal impact of their systems. Even though this regulation is one of the most significant of its kind concerning high-stakes AI systems, it does not provide any definition of "explanation" at any point in the text.

Whereas the GDPR contains no explicit requirement for justification, the AI Act uses the language of "accountability", a legal term closely related to justification. In legal theory, to justify a decision often entails showing that it abides by established standards — precisely the kind of reasoning that underlies accountability in law [27]. Yet, despite this conceptual proximity, the AI Act notably refrains from using the term "justification" itself or from framing obligations in those terms.

This raises the question: why does the legislation stop short of requiring justification, despite acknowledging the importance of accountability? The absence of explicit justification requirements suggests a regulatory gap, or at least an ambiguity, in how algorithmic decisions are expected to meet standards of legitimacy.

To address this ambiguity, we turn to the propositions of the academic literature to clarify how the term "explainability" and related notions are used. As a brief historical note, XAI (acronym for eXplainable Artificial Intelligence) was introduced by the Defence Advanced Research Projects Agency (DARPA) in 2016, which called for innovative research proposals in AI around techniques producing "interpretable" systems that could be understood and trusted by human beings [31]. The term "XAI" has since been widely used to refer to models that provide "insights" into how they arrive at their outcomes [5]. The terms "interpretability" and "explainability" are also widely used in the literature, sometimes interchangeably, to describe the goal of making systems' decisions clearer to human users. By contrast, some authors argue that the two terms should be distinguished, using the term *interpretability* to refer to the availability of a human-understandable description of the system's internals, while anchoring "explainability" in the notion of "explanation", understood as a decision maker/user interface providing the later with details about the model functioning [17, 26].

In a review of opportunities and risks related to the use of algorithmic decision systems (ADS) commissioned by the European Parliament, Castellucia et al. [13] make a valuable effort to clarify such terminologies. ADS are defined as computational systems that assist (or replace) human decision-making by processing large amounts of data to infer correlations. This report is particularly relevant to our study because it articulates a structured set of desiderata. The definition they advance can be summarized as follows:

- "understandability" is the possibility to provide understandable information (for a human being) about the link between the input and the output of the ADS.
- "transparency" is a form of understandability that implies access to information about the internal workings of the algorithm (code, documentation, training data-

sets...). It stands in contrast to "opacity", which characterizes black-box models – systems whose internal workings are either too complex to interpret due to a high number of parameters or deliberately hidden.

- "explainability" is a form of understandability that is defined as the availability of explanations about the ADS. In contrast to transparency, explainability requires the delivery of meaningful information beyond the ADS's descriptive artifacts.

  Therefore, understandability refers to the potential to provide comprehensible information about the system, whether in the form of raw information (e.g., code, documentation) as in transparency, or through the availability of meaningful explanations as in explainability.

- "explanations" are defined technically, either by their form (decision trees, histograms...) or by their type: either operational (informing about how the system actually works), logical (informing about the logical relationships between inputs and results) or causal (informing about the causes for the results), global (about the whole algorithm) or local (about specific results). Thus the availability of any kind of explanations makes the ADS "explainable".

- "interpretability": Castelluccia et al. [13] do not offer any consistently formalized definition of interpretability, as they consider it to be very closely related to explainability. They adopt Lipton's [50] formulation, which states that "explanation is post-hoc interpretability". In general, a system is described as interpretable either when it relies on a simple, transparent algorithm that can be readily understood, or when it produces explanations that render its decisions comprehensible. In that sense, "interpretability" is a synonym of understandability.

- "accountability" refers to the obligation for an agent or system to specify justifications for its decisions, along with the possibility of facing sanctions if those justifications are deemed inadequate. This definition is based on the work of Binns [8] that frames accountability through the democratic political ideal *public reason* as defined by Rawls and Habermas [65, 34]. According to this view, justifications should be based on epistemic and normative standards that can be reasonably accepted by all members of a democratic society.

In the above definitional framework, explanations, of different types, hence serve primarily to enhance understandability. Justification, by contrast, becomes relevant specifically in the context of accountability. In what follows, we expose some typologies of explanations and justifications that can be found in the literature.

2.2 Existing typologies of explanation in the literature

There exist numerous types of explanation methods extensively discussed in the literature, ranging from interpretability techniques for white-box models to post-hoc

approximations for black-box models (see [5, 11, 30, 32, 59]). For space reasons, we leave aside the vast philosophical literature on the classical explanation/understanding dichotomy [72, 67], because it refers to a historically situated debate in the epistemology of historical disciplines, which falls beyond our scope ; although its topic is to some degree akin to our own subject matter, the even vaster philosophical literature on the reason/cause dichotomy [52] is left aside for the same reason.

Hence, we do not claim to survey the literature exhaustively. We rather illustrate the main typologies found in the literature, each defined by different aspects such as what they cover, how they are structured, who they are meant for, and what kind of thinking they use. We introduce a running example to illustrate key features of the various typologies and key-aspects on which they differ.

Example of ADS: a credit scoring system

Let us consider a credit risk assessment system that estimates the probability of default for each loan application. This probability is computed by a predictive model trained on historical data of past borrowers. The model takes as input features such as income level, employment status, existing debt, and credit history, and outputs a value between 0 and 1 representing the estimated likelihood that the applicant will fail to repay the loan (therefore 0 is preferred to 1). Typically, the decision-maker (DM) uses this probability to make a binary decision: if it exceeds a predefined threshold (let's note it t), the application is rejected; otherwise, it is approved.

2.2.1 Types of explanations sorted by scope

The scope of an explanation (or *scale* [57]) is the level at which the explanation is provided. The most widespread categorization distinguishes *local* vs. *global* explanations. Local explanations focus on individual recommendations, while global explanations account for the entire system behavior.

A local explanation targets a **single prediction** $f(x)$. Usually, we aim to identify the *main contributing features* that led to the model's output for a specific instance x. For instance, feature attribution methods such as SHAP (SHapley Additive exPlanations) [53] decompose the prediction:
$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i$ where

- ϕ_0 is the model's base value (the expected output across the dataset),
- ϕ_i is the SHAP value representing the contribution of feature i to the prediction $f(x)$.

Let us consider an applicant described by:

$$x = \{\text{income = 900}\,,\ \text{job} = \text{stable},\ \text{recent\_defaults} = 2\}$$

The model outputs $f(x) = s$ such as $s > t$.
SHAP values for the top features:

$$\begin{aligned} \phi_{\text{income}} &= 0.2 \\ \phi_{\text{recent\_defaults}} &= 0.3 \\ \phi_{\text{job}} &= -0.1 \\ \phi_0 &= 0.4 \quad \text{(base approval score)} \end{aligned}$$

Final score:

$$f(x) = 0.4 + 0.2 + 0.3 - 0.1 = 0.8 \Rightarrow \text{denied}$$

This explanation shows that the decision was mainly driven by *recent defaults* and *low income*, which had stronger negative influence than the positive effect of having a stable job.

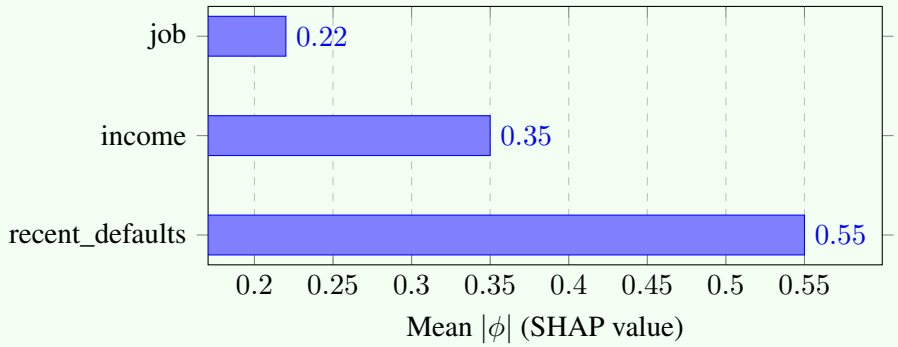
A global explanation targets the **entire model** and aims to provide insight into its overall behavior and logic across all inputs. For instance, one can measure how much each feature generally influences the model's predictions, a measure commonly referred to as *feature importance*. For example, this can be estimated by aggregating the absolute SHAP values across all data samples. Let:

- $\phi_j^{(i)}$ be the SHAP value of feature j for instance i,
- n be the number of data points in the validation set.

The global importance of feature j is computed as:

$$\text{Importance}(j) = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right|$$

A higher average absolute value indicates that feature j consistently contributes more (positively or negatively) to the model's predictions across the dataset.
This method enables ranking features by their overall contribution to the model's behavior and is often visualized as a bar plot as below:



2.2.2 Types of explanations sorted by temporality

Explanations can be categorized according to when in the decision-making process they are generated: *by-design* or *post-hoc*. *By-design* explanations (also called "intrinsic" or "transparent explanations") are inherent to the system's architecture and available during development. These explanations arise naturally from models with interpretable structures, such as some decision trees or linear regressions. In contrast, *post-hoc* explanations are generated after model training to interpret already-made decisions. They attempt to explain black-box models after the fact, often by approximating complex models with simpler, more interpretable ones or by analyzing input-output relationships. This temporal distinction has significant implications for explanation fidelity, as by-design explanations typically offer more accurate representations of the actual decision process. Some of the works that discussed this dimension include [1, 30, 50, 68].

Post-hoc explanation: Suppose the original credit scoring system is a black-box model (e.g., a neural network). Due to its complexity and the large number of parameters it relies on, such a model is not inherently interpretable. To explain a specific decision made by this system, we can use a *local surrogate model* — a simpler model trained to approximate the black-box model's behavior in the vicinity of the instance of interest.
This approach is known as **LIME** (Local Interpretable Model-agnostic Explanations) [66]. It works by generating slight perturbations of the original input instance and querying the black-box model for predictions on these perturbed samples. A simple, interpretable model (typically a sparse linear model) is then fitted to these samples.
For example, consider an applicant described by:

$$x = \{\text{income} = 1000,\ \text{age} = 30,\ \text{recent\_defaults} = 2,\ \text{credit\_history} = \text{poor}\}$$

The black-box model outputs: $f(x) = s$ such that $s > t$.
LIME fits the following local linear model:

$$\text{Local Score}(x) = 0.7 \cdot \text{recent\_defaults} - 0.3 \cdot \text{income} + 0.2 \cdot \text{age}$$

This suggests that the loan was denied primarily due to the applicant's history of defaults. However, it is important to note that this explanation reflects the behavior of the surrogate model in a local neighborhood around x, and not necessarily the global logic of the original black-box model.

Explanation by design: The model is built to be interpretable from the start. For example, consider a sparse linear model:

$$\text{Score}(x) = -0.2 \cdot \text{income} - 0.3 \cdot \text{credit\_history} + 0.5 \cdot \text{recent\_defaults}$$

Each coefficient reflects the contribution of a feature. A higher income or good credit history decreases the score, while more defaults increase it. If the score exceed the threshold, the loan is denied. This model is globally interpretable without the need for external tools.

2.2.3 By target audience

Explanations can be distinguished by their intended audience. We may refer to *popularized* explanations when they are intended for lay recipients, and to *expert* explanations when they target technical or professional stakeholders. The two serve distinct purposes: expert explanation focus on model debugging, validation, refinement, and scientific understanding and *popularized* explanations aim to justify decisions or build trust [7, 66].

Popularized explanations are tailored for non-technical stakeholders, using simplified concepts and accessible terminology. These explanations prioritize understandability and actionability (information that enables someone to take specific, concrete actions based on that information) over technical precision (see for instance the work of Lerouge et al. [49]) As Wick et al. argue [79], these explanations must be reconstructed from technical knowledge into forms that align with recipients' mental models and prior knowledge.

By contrast, *expert* explanations contain technical details and formal representations designed for developers or domain specialists who possess the necessary background knowledge to make sense of them. These explanations may include mathematical formulations, feature importance distributions, model architecture specifications, and performance metrics that would be inaccessible to lay audiences.

The distinction between audience types influences not only the content and form of explanations but also their evaluation criteria. While *popularized* explanations might be assessed on comprehensibility, persuasiveness, and usefulness for decision-making [45], expert explanations are typically evaluated on technical accuracy, completeness, and diagnostic utility [40]. This audience-centric perspective acknowledges that explanation effectiveness is inherently contextual and dependent on the recipient's goals, capabilities, and prior knowledge.

Popularized **explanation:** It should avoid technical terms or parameter values (such as weights in a linear model), even when the underlying model is relatively simple. Presenting raw numerical values can confuse rather than clarify. Instead, explanations should be framed in natural, accessible language that focuses on the user's situation. For example:

$e =$ *"Your loan application was denied mainly because your recent payment history includes a missed credit card payment and your current income is below the required threshold."*

However, such an explanation may be insufficiently persuasive, as users may not understand who defined the threshold, or why that particular value was chosen. A more effective and actionable explanation could be:

$e' =$ *"For the actual amount of credit you are asking for, you should increase your monthly income by 100 euros to be granted the loan".*

This second explanation is more helpful from the user's perspective, as it provides a clear and achievable recommendation to potentially change the outcome.

Expert explanation: In this context, more technical detail is appropriate and often necessary. For instance an explanation for auditors:

$e =$ *"The application was denied because the applicant's risk score fell below the minimum threshold defined by regulatory guidelines. In particular, the requested credit amount exceeded 33% of the applicant's total indebtedness."*

Another explanation for model developers to understand the decision can run as follows. Suppose the ADS is a logistic regression trained on standardized features. The current decision corresponds to an instance x with the following attributes:

$x = \{\text{income} = 950,\ \text{age} = 28,\ \text{credit\_history} = \text{poor},\ \text{recent\_defaults} = 2\}$

The model outputs:

$$f(x) = \text{deny} \quad \text{with probability } 0.84$$

Feature contributions to the log-odds score are: income: -0.45, age: -0.10, credit_history: $+0.60$, and recent_defaults: $+0.80$.

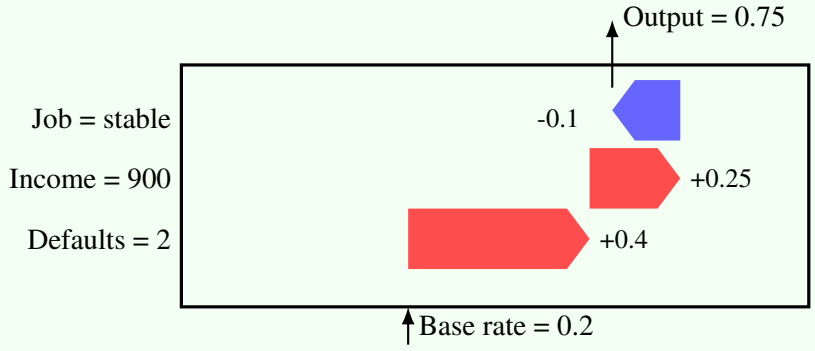
e = "The most influential positive factor toward denial is `recent_defaults`, followed by `credit_history`. Model calibration and AUC = 0.91 indicate good performance, but false negative rate remains high for applicants under 30 with low but stable income."

2.2.4 By format

Explanations can be categorized by their representational format: *visual* or *textual*. Visual explanations include graphs, heatmaps, attention maps, decision trees, feature importance plots, or other graphical representations designed to convey information intuitively. For instance, saliency maps highlighting influential image regions [70] and feature attribution visualizations [53] allow users to literally "see" what the model is focusing on.

Textual explanations, by contrast, rely on natural language descriptions, logical rules, or formal specifications to communicate the reasoning behind model decisions. These explanations range from simple feature-value statements (e.g., "Loan denied due to insufficient income") to elaborate narratives describing causal chains and reasoning steps. Textual explanations benefit from the precision and expressiveness of language, allowing for nuanced descriptions of model behavior and contextual factors. Rule-based textual explanations offer the additional advantage of being machine-actionable, enabling automated verification and reasoning [46].

Visual explanation: The explanation below is a waterfall of SHAP [53] that explains one single prediction. It decomposes the final prediction into the cumulative contribution of each feature, starting from a baseline value (e.g., the average model output) and showing how each feature pushes the score higher or lower. Positive contributions (in red) increase the predicted risk, while negative contributions (in blue) reduce it.



Rule-based textual explanation:
$e =$ "*income* $< 1000 \wedge$ *recent_defaults* $> 1 \implies$ *score* $> 0.6 \implies$ *deny* "
This form uses a logical rule either derived from the model's structure, or approximated from its input-output behavior using techniques such as rule induction.

Natural language narrative: $e =$ *"Your loan application was denied because your reported monthly income of 900 euros is below the minimum required threshold of 1200 euros, and your credit history indicates two recent defaults. These factors suggest a high risk of non-repayment, which exceeds our institution's acceptable risk level."*

2.2.5 By formal reasoning

When generating an explanation, we are implicitly answering a specific type of question such as "Why did this happen?", "Why outcome P instead of Q?", or "What would happen if...?". Each of these questions reflects a distinct form of underlying reasoning, which guides how the explanation is constructed and interpreted. Thus, one meaningful way to categorize explanations is by the type of formal reasoning they embody.

Contrastive explanations address why-not questions of the form "why P rather than Q," where P is the actual event that occurred and Q is a *counterfactual* contrast case that did not occur (a hypothetical alternative outcome). As Miller [55] demonstrates, this form of reasoning is particularly powerful because humans naturally think in terms of contrasts rather than absolute explanations. These explanations focus on the differences between the actual outcome and the expected or desired alternative, highlighting causal factors that differentiate P from Q.

Counterfactual explanations respond to how-to questions and present hypothetical *alternative inputs* that would lead to different outcomes [44, 74]. Unlike con-

trastive explanations, which focus on the outcome, counterfactuals emphasize manipulable features or conditions that would alter the result. These explanations are particularly useful in actionable scenarios where stakeholders need guidance on how to achieve different results. For example, in a loan application scenario, a counterfactual might specify the minimum income needed for approval, whereas a contrastive explanation would explain why the application was rejected when compared to successful applications.

Abductive reasoning forms another foundation for explanations, representing inference to the best explanation [36, 48]. This approach begins with observations and works backward to determine the most likely cause or explanation that accounts for those observations. Abductive reasoning is particularly valuable in complex systems where multiple factors may contribute to an outcome, and the most plausible explanation must be identified from competing possibilities. This form of reasoning often involves evaluating multiple hypotheses against criteria like simplicity, coherence with background knowledge, and explanatory power.

Causal reasoning underlies explanations that identify genuine cause-effect relationships rather than mere correlations. These explanations answer what-if questions by mapping the causal structure that connects inputs to outcomes. Drawing from Pearl's causal calculus [62], causal explanations distinguish between direct effects, indirect effects, and spurious associations. This type of reasoning allows for more robust explanations that can anticipate system behavior under novel conditions or interventions, making it particularly valuable for critical applications where understanding the underlying mechanisms is essential.

Contrastive explanation: *"Your loan application was denied (P) rather than approved (Q) because your reported income was 1000, whereas applicants who were approved had incomes above 1500, and you have two recent defaults while successful applicants had none."*

Counterfactual explanation: *"If your monthly income had been at least 1200, and you had no defaults in the past year, your loan would have been approved."*

Abductive explanation: *"The loan was denied likely because of a combination of low income, multiple recent defaults, and poor credit history. Among these, recent defaults appear to be the strongest contributing factor."*

Causal explanation: *"Your loan was denied because your income directly influences the risk score, and income below 1200 causes the score to fall below the approval threshold. Other factors like age and employment have indirect*

or negligible effects in this case."

2.2.6 By discursive interaction

Another body of work move away from viewing explanations as static elements and instead present them as *processes* that unfold through interaction. As Miller [55] argues, explanations should be understood as forms of social interaction rather than as static pieces of information. In this perspective, an explanation is not merely a statement of causes or reasons but a *conversation* between an explainer and an explainee. Drawing on Hilton's [39] conversational model of explanation, Miller highlights that explanations, like conversations, are governed by cooperative principles. They must address the explainee's question, remain relevant to the context, and adhere to Grice's conversational maxims of quality, quantity, relation, and manner [29]. These maxims imply that a good explanation should be truthful and supported by evidence (quality), sufficiently informative without unnecessary detail (quantity), pertinent to the recipient's concern (relation), and expressed clearly and coherently (manner).

Closely related to the dialogue perspective, the dialectical approach adds a layer of *argumentation*. While dialogue aims at mutual understanding through cooperative exchange, dialectic seeks to challenge and justify positions through reasoned debate. It thus goes beyond a simple question–answer model, aiming not merely to report causes but to support claims with arguments [76, 78]. Walton [78] propose dialogical models to implement Hilton's conversational model, composed of three stages: an opening stage (where the need for an explanation is established), an exploration stage (where reasons are exchanged), and a closing stage (where understanding is assessed). This model addresses a central issue: how to determine when an explanation is complete. Walton [76] argues that understanding is not achieved merely when information is transmitted, but when the explainee can demonstrate comprehension by answering new, related questions. Moreover, selecting the *best* explanation does not mean the best one possible (as in abductive reasoning), but rather identifying the explanation that is arguably stronger than its alternatives. Because explanations rely on *defeasible* reasoning —that is, reasoning open to revision or challenge — they are naturally framed within argumentation schemes such as arguments from analogy, from expert opinion and so on [77].

In a socioeconomic perspective, scholars from the "theory of conventions" put the emphasis on the fact that in the economic and social world, there is a variety of frameworks of interactions, named by Boltanski and Thévenot [9] the "cités". For instance, in the "market cité", the motive for action is the profit. In the "industrial cité", the cooperation between workers, bosses and machines is guided by the search for productive efficiency. In the "civic cité", what is valued are the democratic principles or the social commitment of citizens. To sum-up what appears valuable in this theory,

in each cité a supreme principle provides a shared communicative pattern based both on reciprocal expectations. The common supreme principle is not intangible: it is continuously subject to criticism. Boltanski and Thévenot call this process "justification". The principles of profit in the market, of technical efficiency in the industry, of social commitment in the community can be criticized and must prove their legitimacy. From this socio-economic theory, we must retain the ideas that the justification operation presupposes a common framework, and that it must be appropriate to the context in which it is deployed.

Hénin and Le Métayer [38] also emphasise the importance of justification and contestability. They provide clear definitions of each concept and explicitly distinguish between explanation and justification, as we also aim to do in this paper (in subsection 3.1, we further clarify how our approach differs from theirs). Their work is also situated within this typology of explanatory models, since contestability introduces a dialectical dimension and they do it with reference to norms. In their framework, called *Algocate* [37], users can interact directly with the decision system by asking questions and receiving tailored explanations, thereby transforming explanation into a cooperative conversational process. The third example below is inspired by their framework.

Dialogical explanation Suppose an applicant's loan request is denied and the decision-maker provides an interactive interface that allows the applicant to ask questions and receive contextually adapted explanations. This interactive approach operationalizes the conversational nature of explanation proposed by Hilton [39].

User "Why was my loan denied, even though my revenue is greater than 5000 and my credit amount is lower than 2000"

System "Your recent credit history shows two missed payments, which increased your estimated risk of default."

User "What can I change to be accepted?"

System "If your average income over the last six months were 200 higher, your risk score would fall below the approval threshold."

Dialectical explanation Here, the user can challenge the reasons offered, question the decision's legitimacy and the model's assumptions. Explanation thus evolves into justification through argumentation, reflecting a shift from understanding to critical deliberation.

User "Why was my loan denied even though my income is stable?"

System "Your loan application was denied because, according to financial experts, applicants with more than two recent defaults are highly likely to default again. Our system follows this principle to minimize risk." (*argument by expert opinion*)

User "I understand, but my defaults were due to a short-term health issue. Since then, I have repaid all my debts on time. Shouldn't my recent repayment history count more than older defaults?"

System "The rule we apply generalizes from expert knowledge, but your recent financial stability may indicate that this generalization does not hold in your case. We can review your situation in light of this new evidence."

Norm-Based justification in the sense of Hénin and Le Métayer: suppose an applicant contests the decision so the decision-maker has to justify it.

DM Your loan application was denied because your debt-to-income ratio exceeds 45%. According to our institution's credit policy, based on the national financial regulation "Code du crédit à la consommation", loans cannot be granted above this threshold.

User But I've recently obtained a permanent job that increases my income. Shouldn't that be taken into account before applying the rule?

DM The regulation aims to prevent over-indebtedness and ensure borrowers' financial stability. Even though your situation may improve, we must apply the same standard to all applicants to remain compliant with this fairness principle.

This dialogue illustrates a *norm-based justification*, in which the decision-maker refers explicitly to a regulatory norm to legitimize the outcome. In contrast to mere *explanation*, which describes causal mechanisms, *justification* appeals to shared principles (legality or procedural consistency). The exchange also reflects Hénin and Le Métayer's idea that justification connects the decision to the normative framework governing the system.

To conclude, figure 1 summarizes the conceptual evolution of explanatory approaches in the literature. Early work focused primarily on the *structural* dimension of explanation, emphasizing model interpretability through mechanisms such as feature importance or visualization. This was followed by *cognitive* approaches, which aim to align explanations with human reasoning processes and cognitive biases, as discussed by Miller [55]. A further shift introduced *epistemic* perspectives, empha-

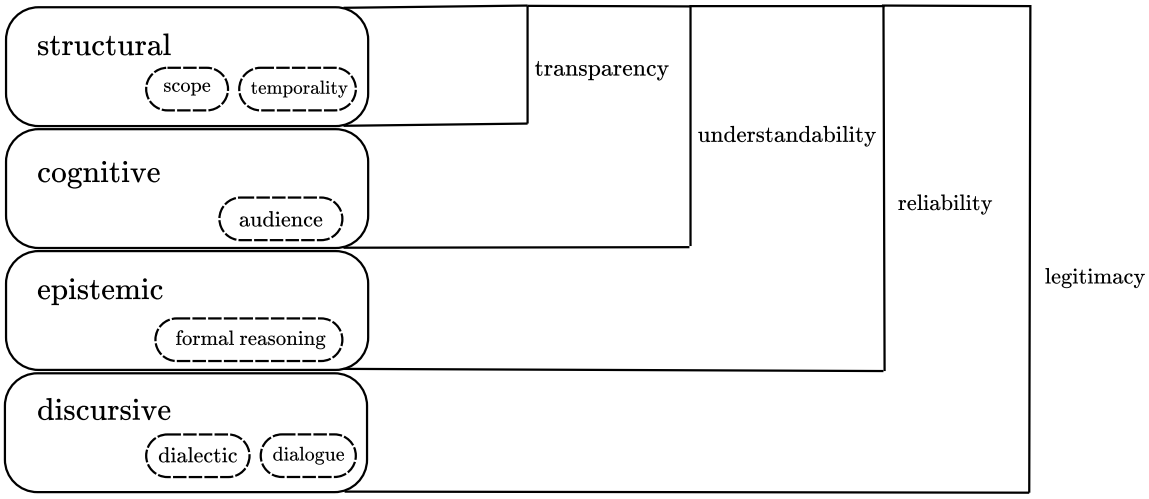


Figure 1: Conceptual bridge between explanation typologies

sizing the explanatory value of causal knowledge and truthfulness in model reasoning. Finally, other work extends toward *normative* approaches, which view explanation as a discursive process grounded in social interaction and argumentation. These notions can be understood as forming a hierarchy of inclusion. Transparency represents the most basic level, alone it is insufficient if the information provided cannot be meaningfully interpreted, hence the need for comprehension, which presupposes transparency but adds information that aligns with the expectations and mental models of the recipients. Reliability builds upon comprehension, since users can only develop trust in a system they understand and perceive as consistent in its behavior. Finally, legitimacy encompasses all previous levels: a system can be transparent, understandable, and reliable, yet still fail to be legitimate if its outcomes are not perceived as such or aligned with shared norms.

Our work is situated within the discursive category, and builds upon it. While prior argumentative approaches succeed in capturing the interactional dimension of explanation, we argue that they do not ensure *acceptability*. Achieving *acceptability* requires more than legitimate process: it requires that the decisions be justified in the eyes of their recipients. To articulate this distinction precisely, we draw on Habermas's theory of communicative action, which provides a framework for evaluating the rationality of communicative exchanges, and on Perelman's theory of argumentation, which grounds the success of justification in the recipient's reasoned adhesion. On this basis, we propose a clear conceptual distinction between *explanation* and *justification*. This distinction is not merely terminological: it identifies a gap in existing approaches and motivates a new conceptual framework, developed in Section 3, designed to bring algorithmic decision-making systems closer to collect acceptability.

3 A conceptual framework

The former section explored the vast, multidisciplinary literature on explanation, explainability and justification. This review highlighted the fact that this literature presents different typologies of concepts that are all to some extent relevant, but stem from different logics and hence explore different aspects of the diversity of concepts, issues and ideas surrounding explanation, explainability and justification. Taken one by one, the different typologies appear focus on specific aspects of their subject matter, but an overarching conceptual framework encompassing the specific strengths of these various approaches and building upon their complementarities is found lacking in the literature. Our point in this section is to propose such a conceptual framework.

The crux of our proposal is to conceptualize explanations and justifications as unified by their shared anchorage in argumentation and social exchanges mediated by language, but distinguished by the specific roles they play in different kinds of discussions with different types of recipients.

The shared anchorage we propose refers to Meinard and Tsoukiàs [54], who argue that the philosophy of Jürgen Habermas can usefully shed light on the analysis of forms of rationality in decision support theory and practice. They introduce an analytical framework leading to a Habermas-inspired typology of decision aiding approaches. Their typology, which distinguishes between "objectivist, conformist, adjustive and reflexive approaches", is intended to be useful to practitioners, who should be able to identify which approach they should use in a given situation. Below, we show how this framework can be used as a common frame for both explanation and justification, and how it manages to encompass the strengths of the various typologies explored in the former section.

Within this common frame, we propose the following clarification of the distinctive roles of explanation and justification. Whereas explanations serve to clarify the decision-making process, justifications are designed to ensure that the decision will be perceived as convincing and legitimate. This aligns with Perelman's concept of argumentation, where the success of the exchange is determined by its capacity to gain the recipient's acceptance, which is achieved through a combination of clear explanations and persuasive justifications [63].

For the purpose of developing and presenting this conceptual proposal, we make some working simplifications. Although automated decision-making systems usually involve multiple stakeholders, with multifarious roles, rights, duties and stakes, such as the decision-maker, the system-designer, the lay recipient of the decision, the explanation and/or justification issuer, the recipient of such explanation and/or justification...

In what follows, we essentially focus on who issue the explanation and/or the justification (denoted IS) and who receive them (denoted RCP) since our framework is built upon their interaction. As a further simplification, regarding the algorithm,

we do not differentiate between white-box and black-box models, as our focus is on the presence of an algorithm —regardless of its nature and internal complexity—in the decision-making process. In both cases, the need for explainability remains the same. However, the type of explanation required differs. White-box models may offer inherently interpretable explanations due to their simple structure, whereas black-box models often require post-hoc methods to provide meaningful insights into their decision-making process.

3.1 Explanation vs. Justification

Let us define an **explanation** as the operation that consists in giving the characteristics of the process that led to the decision. In other words, an explanation consists in informing the RCP (whether it is the decision-maker or the recipient of the decision) about what actually happened in the decision-making process, factually and historically. The main goal of an explanation is that its recipient should *understand*.

As opposed to explanations, a **justification** goes beyond factual recounting (the hallmark of explanations) to assess whether the decision was normatively acceptable. Justifications offer reasons aimed at establishing the legitimacy of a decision, meaning reasons that the affected individual can find reasonable.

Both explanations and justifications are shaped by practice, as they primarily concern an action or a disposition to act [64], and they are both typically context-dependent. Because the context is not always immutable, various elements of both explanations and justifications can be subject to reconsideration if compelling reasons are presented to challenge them [64]. In Habermasian terms, both explanations and justifications are situated in the "life-world" [33], which implies that they must resonate with people's everyday understanding and shared social norms. In justifications, this reliance on everyday norms tends to overweigh the reliance on technical details, while the later can play a more important role in explanations. However, both elements can have a role to play in both explanations and justifications.

Other scholars have already given significant weight to the difference between explanation and justification. For example, for Hénin and Le Métayer [38], the distinction between these notions is relevant because they have specific properties and goals. They consider "Explanations (as) transfers of knowledge (from the ADS to the explainee)" that are descriptive and intrinsic; in contrast, justifications are normative and extrinsic in the sense that the judgment about the goodness of an outcome must be based on an external reference such as a legal or ethical norm. Although we concur with Hénin and Le Métayer that explanations and justifications should be distinguished, we depart from them on two important ideas. First, the reference to an external norm is not a necessary condition for a justification; second, a justification is connected to a search of adhesion, in an argumentative process. These ideas will be further elaborated in the next sub-section in which the explanation/justification issue

will be framed with reference to a typology of explanation models.

Same context as 2.2. Suppose we have $f(x) = s \geq t$ for individual x, the DM provide an explanation to the applicant :
e_{DM} = *"the loan was denied because the current income is below the threshold required to safely cover monthly loan installments"*.
In this case, if the explanation process ends here, the applicant may question the legitimacy of the threshold and find the explanation unacceptable.
A justification could be:
j_{DM} ="*The decision was made in accordance with the institution's credit policy, which is designed to ensure applicants are not placed at financial risk they cannot bear. The income threshold is set based on statistical models and regulatory guidelines that balance financial inclusion with responsible lending. In cases where applicants are near the threshold, we offer alternative financial products or re-evaluation after updated income data.*"
This constitutes a justification because it provides normative reasoning and offer reasons that a reasonable person could accept, even if they disagree with the outcome.

3.2 A typology of explanation and justification models

We propose a typology which distinguishes four explanatory models. It is build mainly upon Habermas' models of action and, in a specific model, upon Perelman's theory of argumentation. In what follows, the models we propose are *ideal-typus* of explanation and justification of decisions based on an ADM. We will argue that the four models that will be proposed do not perform equally in terms of justification, adherence and finally legitimacy.

3.2.1 Technical model

This first type is based on Habermas' strategic model of action, which accounts for an agent's action in terms of objective facts [54]. Here, the algorithmic model is considered rational owing to its solid scientific and technical grounds. A technical explanation is hence exclusively based on scientific and technical considerations. A technical explanation can be used to explain the functioning of a given algorithm to the decision issuer. If used to convince a lay recipient that this algorithm produced legitimate outcomes, or functions in a legitimate way, then in this usage it plays the role of a justification. Using technical explanations in this way can foreseeably have inconvenient implications:

- It is likely that there will be a gap in understanding between IS and RCP.

- For the RCP, the justification provided can fail to be convincing, and the legitimacy can hence be contested.

3.2.2 Norm-oriented model

In this model, both explanations and justifications are based on compliance with legal or moral standards (such as non-discrimination, equal treatment, mandatory human intervention to issue the final decision...). An explanation is norm-oriented when it accounts for the objective, historical fact that this or that choice was made, when elaborating the algorithm, *in order to* comply of the norm at issue. A norm oriented justification argues that the functioning or the outcome of the algorithmic system is legitimate *because* it complies with this norm, whether this compliance was intended beforehand or not.

The following corollaries are worth mentioning:

- A norm-oriented explanation formalizes and explains the specific norms that were used when elaborating the algorithmic model, their interpretation, and the way they are implemented in the model.
- Legal norms and subjective fairness do not necessarily overlap. Legality or alignment with social norms certainly are sources of legitimacy, but if a given decision is substantially disadvantageous for a given RCP, the general justifiability bestowed by legality or social norms can be outweighed by personal disadvantage in the RCP's overall assessment. In such cases, norm-oriented justifications can require, in order to be successful, detailed and convincing justifications of why a given (generally acceptable) norm has had this or that particular implication for the RCP.
- As opposed to norm-oriented explanations, a norm-oriented justification may appeal to a norm that did not play any role in the design and motivation of the algorithm, if the functioning and/or outcome appear to abide by this norm. This approach can feed more powerful justifications, if the norm at issue is more widely shared and hence proves more convincing for the RCP.

3.2.3 Expressive model

The hallmark of the expressive model is the idea that rationality stems from the sincere expression of motives, preferences and values. An expressive explanation will hence faithfully track and report the motives, preferences and values that the designer of the algorithmic model has (possible unconsciously) expressed through the choices s/he has made when elaborating the algorithmic model. An expressive justification will herald motives, preferences and values that its recipient can recognize as his/her own.

3.2.4 Communicative model

The fourth model echoes Habermas's "communicative rationality", which involves impartial discussions and the search for the best argument. Its implementation implies a communicative activity. The model is also consistent with Perelman's theory of argumentation, which stresses the following ideas: argumentation essentially is a search for convincing argument; it unfolds as a progressive process; it must be coherent with the recipient's needs and expectations; the acceptability of an argument is context-dependent; there is hence no such thing as an argument that would be totally binding in itself.

Explanation and justification, in the communicative model, hence involve a communicative activity that consists in producing and exchanging arguments. When this communicative activity serves the purpose of fine-tuning the account given of how the algorithmic system functions, in order to foster the recipient's understanding, then this communicative activity yields a communicative explanation. If it rather aims at convincing that decisions or recommandations are legitimate or well-founded, then it constitutes a communicative justification.

This model is largely, though not entirely, consistent with Miller's framework [55], which is itself inspired by Hilton's conversational model [39] and Walton's dialogical theory of explanation [78]. Our model shares with these alternative frameworks an emphasis on conversation and dialogue between explainer and recipient.

3.2.5 Summary

Table 1 synthesises the conceptual framework developed above. A key observation is that the same model serves different purposes depending on whether it is used as an explanation or a justification, and that these purposes do not always align.

A *technical* explanation is a correct recount of how the decision was made; a *technical* justification shows how the decision complies with standards of scientific accuracy. A *norm-oriented* explanation accounts for the norms that actually motivated the design of the system; a *norm-oriented* justification reflects more the issuer's policy choices than the norms that were operationalised in practice. An *expressive* explanation faithfully tracks the motives and values expressed through the designer's choices; an *expressive* justification appeals to motives and values the recipient can recognise as their own. Whether these two sets of values coincide is an empirical question, and when they do not, expressive justification will fail.

A *communicative* explanation deploys argumentation to foster the recipient's understanding of how the system functions; a *communicative* justification deploys the same argumentative resources to establish that the decision is legitimate and well-founded. The unifying activity — the exchange and critical assessment of arguments — is the same in both cases. What differs is the goal. Explanation succeeds when

the recipient finds the account acceptable, that is, intelligible, relevant, and consistent with their expectations and mental models. Justification succeeds when the recipient finds the decision legitimate, that is, when the reasons offered can be endorsed through reasoned deliberation, even in the face of a personally unfavourable outcome.

Model	**Explanation**	**Justification**
Technical	Scientific accuracy Correctness	Compliance with technical standards
Norm-oriented	Compliance with existing norms	Policy / rationality oriented
Expressive	Faithfulness	Meeting the recipient value system
Communicative	Argumentation	
	Acceptability	Legitimacy

Table 1: Conceptual framework of explanations and justifications

3.3 Implications of the overall structure of the conceptual framework

By definition, one cannot expect a justification to be successful if its issuer does not share some common ground with its recipient. Sharing the same model appears to be a minimal requirement in this regard. The choice of a model that the recipient will endorse is hence crucial for any issuer of a justification. When making this choice, the issuer should be aware that the communicative model (model 3.2.4) has distinctive strengths when the interests of IS and RCP diverge.

Notice that the justification given by the IS may be strategic: it may indicate reasons deemed acceptable to the RCP, instead of the true reasons. For instance, consider a company that uses an automated system to filter job applicants based on their residential address, systematically rejecting candidates who live outside a specific, affluent neighborhood. On the surface, the IS might offer the following justification: *"We prioritize hiring people who live nearby in order to support the local economy and strengthen community ties."* While this justification may appear socially responsible, it can mask underlying discriminatory motives, such as using geographic location as a proxy to exclude candidates from less wealthy or more ethnically diverse areas. In this case, the justification functions more as a socially acceptable narrative than as a sincere account of the decision's normative legitimacy.

4 Case study: admission in French universities

In this last section, we develop the application of our conceptual framework in a case study: the process used in France for university admission at two levels: undergraduate (first-year) studies and Master's programs. We analyze this empirical case through the lens of the above conceptual framework and use the later to assess the current system in an original perspective. Based on these findings, we draw lessons that might prove useful for both developers and decision-makers in their efforts to improve the way rebuttals are justified to students.

4.1 Properties and functioning of the system

Access to universities in France has undergone radical change since the late 2000s. Traditionally, the right to higher education has been a prominent principle of the French system. Unlike so-called "grandes écoles", where admission is decided by highly selective entrance examinations, admission to university has traditionally been open to all. However, this principle of a "right to higher education" was challenged in recent years by the use of automated algorithms for selecting students applying to different university disciplinary programs. After an initial experiment in the "Ile-de-France" Region in the 1990s, the State introduced a national application in 2009: APB ("Admission Post-Bac"), which was replaced by a new algorithm in 2018, Parcoursup, which is currently still in use.

Under the APB system, high school students in their final year had to prioritize wishes (with a maximum of 24 wishes). Assignments were made using local algorithms determined by the universities. APB was plagued by numerous malfunctions in high-volume disciplines such as medicine, law, psychology and sport, in which lotteries had often to be used to decide between numerous candidates of equal merit. This practice lacked a legal basis, as stressed by the highest French administrative court - the "Conseil d'Etat" - in a decision dated December 22, 2017. The Ministry of National Education has subsequently replaced APB with Parcoursup, which has been legally in force since March 8, 2018 ("loi relative à l'orientation et à la réussite des étudiants", known as "the ORE law"). Parcoursup is based on the "stable matching problem", solved using the Gale-Shapley algorithm [24]. In this process, students in their final year of high school submit 10 non-hierarchical wishes for each academic program, which can be broken down into 10 sub-wishes for each university.

The procedure comprises 3 phases, with a very tight timetable: the platform opens in November, at which point students fill in a "dialogue form" for the Class Council's opinion; from January to March, applicants formulate their wishes; then, from May to July, the universities' responses are communicated, and students make their final choice. Parcoursup concerns access to undergraduate studies. Access to post-graduate courses is orchestrated by a dedicated platform: Monmaster. In principle,

universities must provide as much information as possible about their selection and admission criteria, as well as their intake capacity for 1st year students in the various courses, specifying the number of places reserved for students who benefit from social scholarships. If rejected or placed on a waiting list, candidates have two options. The first option is a hierarchical appeal. Such appeals are handled by the "Commission d'Accès à l'Enseignement Supérieur" (CAES), under the responsibility of the "Recteur" (the regional head of the education administration). There are around 23,000 appeals per year (an apparently stable figure: 23000 in 2018, 25000 in 2019, 23400 in 2023). The alternative option is to submit a discretionary appeal. Such appeals are submitted either to heads of courses (usually by e-mail), and then consist in a request to be admitted to a given first year program or to a given Master's course, or to President of Universities (by registered letter), in which case they inquire about the reasons motivating rebuttals. There are between 500 and 1,000 appeals of that sort per year in a specific department" which "can lead universities to process appeals automatically, generating standardized responses which are calibrated in advance by the legal and academic departments" (Allouch and Espagnolo-Abadie,p. 21 [3]).

Although Parcoursup is a national platform, it is implemented at a local scale by universities, which are autonomous since a reform initiated in 2007. This leads to a wide variety of academic assessment criteria, depending on the university, or even on the department within a given university, which is in charge of applying these criteria. Moreover, the context matters: not all universities operate under the same logic when it comes to student selection. Some institutions have a long-standing elitist culture, and for them, the introduction of student selection via Parcoursup is not a break with the past - quite the contrary. Others are still attached to the right of all "baccalauréat" holders to access higher education; they have reluctantly switched to a selective system and are critical of the selection process, which they claim disadvantage underprivileged social classes. A third group of universities combines low selectivity for first-year admissions with high selectivity at the Master's level. In this context, appeals against first-year non-admissions mostly arise in the second category of institutions, while appeals concerning Master's-level rejections are primarily seen in the third.

The variety of academic assessment criteria mentioned earlier is perceived as a source of opacity by applicants to the Parcoursup algorithm, as underlined by a 2020 parliamentary report [42]. This perceived opacity is reinforced by the lack of information on the selection methods used in practice: some universities select "by hand", while others use local algorithms.

According to a 2023 IFOP (a pollster) survey, 83 percent of respondents find the Parcoursup experience stressful. A 2019 survey by the "Observatoire de la Vie Etudiante" showed that 27 percent of students find the Parcoursup platform unfair [22]. Indeed, appeals against non-admission decisions are based on a feeling of injustice, particularly when students have a very good academic record and adhere (along with

their families) to the meritocratic system, and all the more so when the parents have higher education qualifications, belong to the upper classes and are attentive to their child's studies.

4.2 Assessment: does the system manage to successfully justify rebuttals?

The problem of justification of rebuttals is a sensitive one. Candidates whose demands have been rejected are very often (if not always) given standardized explanations, sometimes automatically generated by local algorithms. When program managers receive 500 to 1,000 requests in a short period, all asking for the reasons behind their rejection, they lack both time and human resources to respond adequately. Moreover, professors are generally not accustomed to this kind of exercise. There clearly is a discrepancy between the expectations of rebutted candidates (and their families) – particularly when applicants were very good high school students, invested in their education.W hen confronted with a non-admission decision in their most preferred academic programs, they feel disoriented or destabilized by both the negative decision itself and the nature of explanations eventually provided. University professors are aware of this tension but often feel have no viable alternative [3].

The uncomfortable position of university professors has been documented by Allouche and Espagnolo-Abadie [3]. They quote the head of a psychology department expressing discomfort about the selection process, arguing that relying simply on grades instead of applicants' motivation is "more practical and more protective for teaching teams". This professor also admits that, given the limited number of available spots, very good candidates may end up being rebutted. These empirical elements can be integrated into the typology of explanation and justification models defined above (in section 3).

- Technical Model: the IS (in this case, the Ministry of Education) has developed a Gale-Shapley algorithm that is supposed to be fairer and technically more advanced than the previous algorithm ("Admission Post-Bac" - APB). Technical assistance is provided for users. However, this technical assistance is not what rebutted applicants expect. Indeed, the later expect assistance in terms of guidance advice, answering questions such as: is it appropriate to include this or that course or university in their wishes? The assistance device fails to provide candidates with relevant answers to such questions, and rather focuses on algorithm design. Therefore, what the assistance provides is not seen as an *explanation* by users, since it does not focus on the aspects of the system's functioning that those users are interested in; *a fortiori* it does not provide a convincing *justification*.

- Norm-oriented model: Parcoursup designers consider the algorithm to be fully compliant with legal norms, including the law that mandated its use (the ORE law), the GDPR and relevant European regulations that French law must comply with. In other words, the IS asserts that the system operates within the bounds of legal norms. However, rejected applicants are not primarily contesting the legality of Parcoursup. Rather, they challenge the gap between the decisions it produces and pervasive meritocratic social norms, namely the belief that a good student deserves access to the university of his or her choice to pursue the studies of their choice. Thus, compliance with legal norms from the IS's perspective may generate a disappointed meritocracy. In this case, two kinds of norms are at odds: legal and social norms. The argument of legality provides a kind of justification which confronts the wish of a meritocratic-based admission in university. Available norm-oriented justifications hence fail because the norms they refer to, which can indeed by the ones that a norm-oriented explanation should refer to, are not the norms that applicants would accept as structuring in convincing justification.
- Expressive model: rejected applicants often criticize the dehumanization of university admissions and the impossibility of direct contact with course managers or university decision makers. Meanwhile, the IS considers the algorithm as a highly efficient matching system between applicants' preferences and academic programs capacities. The reasons given to rejected applicants are standardized and fail to meet their expectations, particularly for those who believe in selection based on merit. Allouche and Espagno-Abadie [3] (p. 110 to 113) cite parents' criticisms of the system: some consider that everyone hides behind the Parcoursup algorithm, others that no convincing explanations are given and even that dehumanization leads to their child's value being discredited . Therefore we can conclude that the provision of explanations about the properties of the algorithm fail to satisfy the demand of justifications expressed by the candidates and their family.
- Communicative model: rejected applicants tend to express two types of expectations: first, to have direct human interaction with academic program coordinators to try to convince them that they deserve to be admitted, and second, the need to understand the reasons behind their rejection. Most of the time, they request both an explanation for the decision and a justification. These expectations are often unmet due to the absence of opportunities for meaningful communication or dialogue with the university. Furthermore, the actual reasons behind admission refusals are frequently concealed. For example, according to Allouche and Espagnolo-Abadie [3], regulations prohibit selecting Master's candidates based on grades. Yet, according to one Master's program director in law cited by the authors [3], selection in practice is indeed grade-based. The

Monmaster platform then records a "satisfactory level", which is supposed to "justify" the refusal, when those admitted have received an assessment of "excellent level". The two authors cite the case of a student applying for a Master 2 in notarial law. She had a very respectable academic record, but was rejected. She felt that her career plan, which was to become a notary, her motivations and the fact that she had completed internships in the profession, had not been taken into account. In other terms, the communicative model is able to provide a convincing justification only in the case where a genuine dialogue takes place between the decision issuer and the recipient. All in all, the system as it is currently used hences does not produced any communicative explanation or justification.

The analysis above shows that, although the system as it stands involves some attempts at explaining and justifying decisions made, by and large, these attempts fail. As highlighted in our discussion of explanations anchored in the technical model, such tentative explanations fail because they fail to grasp the expectations of their recipients. Similarly, as shown in our discussion of norm-oriented attempted justifications, available justifications misidentify the norms that recipients are liable to adhere to, and tend to mistake a norm-oriented explanation for a norm-oriented justification. A thorough application of our framework in the very process of producing explanations and justifications would help avoid such mistakes, and thereby help design more satisfactory explanations and more convincing justifications.

That being said, as highlighted in our discussion of an absence of any communicative approach in the system as it stands, because the decisions made through this algorithm are highly sensitive and because the context is very complex in most cases, it seems clear that applications of the communicational model are the most promising. The challenge, for designers and champions of this system, is hence to develop technical means to enable an argumentative dialog between issuers and recipients of explanations and justifications.

To be sure, such perspectives are bound to be hampered by the increasing shortage of human and material resources in French universities. But that is another story.

5 Conclusions

Contemporary societies are unmistakably characterized by a growing pervasiveness of decision-support and recommendation systems, mostly algorithmic in their structure, spanning all the domains of our ordinary live (from everyday consumption to the organization of professional live, through leisure activities, medical care and education) as well as exceptional events such as natural disasters or financial crises.

The pervasive use and very existence of these algorithmic systems have profound,

and often unintended and/or unanticipated effects on the states and dynamics of numerous aspects of our societies, including inequalities in income, life conditions and life prospects of various people and communities. Accordingly, there is a growing awareness that these systems should abide by certain norms and expectations, which have historically been applied to human agent rather than to algorithmic systems, and therefore require being adapted to these objects.

Exploring the literature addressing this pressing need, we have shown that existing legislative and academic efforts, through useful in providing diverse relevant insights to think through these issues, lack an overall, general framework accounting for these various aspects in a convincing manner.

In order to bridge this gap, we have introduced such a general framework, inspired by fundamental contributions in the philosophy of social sciences [33] and methodology of decision support [54]. The crux of our proposed general framework is a two-fold clarification:

- a first clarification consists in distinguishing, on the one hand, *explanations*, understood as accounts of the *actual proceedings of the process* and, on the other hand, *justifications*, defined as *convincing and legitimate reasons why the process or its outcomes should be considered acceptable*.
- a second clarification consists in distinguishing *four models (technical, norm-oriented, expressive and communicative) of both explanations and justifications*, anchored in Habermas's typology of models of action.

A focal case study, the algorithmic decision process used to adjudicate admissions to French universities, was used to illustrate the usefulness and added-value of our proposed framework. Further attempts at exploring the promises and limits of our proposed framework are now needed to entrench its usefulness and/or refine its structure.

The article and works on the legal constraints of explainability of algorithmic decisions or their feasibility are complementary. Our focus here was not on legal regulation *per se*. Our aim was rather to provide a framework of analysis that can shed light on legal norms concerning requirements articulated in terms of explainability and associated phrases.

6 Acknowledgments

The authors thank the Mission pour l'Interdisciplinarité et les Initatives Transervses of the CNRS (Paris) for financial support to the Fairness by Explanation of Algorithmic Decision (FEAD) and SPLEAD projects.

References

[1] Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access **6**, 52138–52160 (2018)

[2] Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., Gao, G.: Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. Health Policy and Technology **12**(1), 100702 (2023)

[3] Allouch, A., Espagnolo-Abadie, D.: Contester Parcoursup. Presses de Sciences Po (2024)

[4] Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society **20**(3), 973–989 (2018)

[5] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion **58**, 82–115 (2020)

[6] Bates, T.: Technology and culture: How predictive policing harmfully profiles marginalized people groups. In: California Sociology Forum. vol. 6, pp. 18–27 (2024)

[7] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M., Eckersley, P.: Explainable machine learning in deployment. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 648–657 (2020)

[8] Binns, R.: Algorithmic accountability and public reason. Philosophy & technology **31**(4), 543–556 (2018)

[9] Boltanski, L., Thévenot, L.: On justification: Economies of worth. Princeton University Press (2006)

[10] Bono, T., Croxson, K., Giles, A.: Algorithmic fairness in credit scoring. Oxford Review of Economic Policy **37**(3), 585–617 (2021)

[11] Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (2021)

[12] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)

[13] Castelluccia, C., Le Métayer, D.: Understanding algorithmic decision-making: Opportunities and challenges. Tech. rep., European Parliament (2019)

[14] Cinus, F., Minici, M., Monti, C., Bonchi, F.: The effect of people recommenders on echo chambers and polarization. Proceedings of the International AAAI Conference on Web and Social Media **16**(1), 90–101 (May 2022)

[15] Colorni, A., Tsoukiàs, A.: What is a decision problem? European Journal of Operational Research **314**, 255 – 267 (2024)

[16] Danaher, J.: The threat of algocracy: Reality, resistance and accommodation. Philosophy & technology **29**(3), 245–268 (2016)

[17] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

[18] Edwards, L., Veale, M.: Slave to the algorithm? why a "right to an explanation" is probably not the remedy you are looking for. Duke Law & Technology Review **16**, 18–84 (2018)

[19] Eubanks, V.: Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press (2018)

[20] European Commission: EU AI Act. Council of the EU, Press release (December 2023), https://artificialintelligenceact.eu/ai-act-explorer/

[21] European Parliament: General Data Protection Regulation. Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (May 2016), http://data.europa.eu/eli/reg/2016/679/oj

[22] Feres, B., Huilton, C., Odile, F., Élise, T.: L'orientation étudiante à l'heure de parcours sup. des stratégies et des jugements socialement différenciés. OVE Info **39** (2019)

[23] Fourcade, M., Healy, K.: Classification situations: Life-chances in the neoliberal era. Accounting, Organizations and Society **38**(8), 559–572 (2013)

[24] Gale, D., Shapley, L.: College admissions and the stability of marriage. The American Mathematical Monthly **69**, 9–15 (1962)

[25] Gillespie, T.: Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press (2018)

[26] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89. IEEE (2018)

[27] Golltzberg, S.: Chaïm Perelman.L'argumentation juridique. Michalon - Le bien commun (2013)

[28] Green, B., Chen, Y.: The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction **3**(CSCW), 1–24 (2019)

[29] Grice, H.P.: Logic and conversation. Syntax and semantics **3**, 43–58 (1975)

[30] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)

[31] Gunning, D.: Explainable artificial intelligence (XAI). darpa-baa-16-53, Defense Advanced Research Projects Agency (2016)

[32] Gunning, D., Aha, D.: DARPA's explainable artificial intelligence (XAI) program. AI magazine **40**(2), 44–58 (2019)

[33] Habermas, J.: The theory of communicative action, vol. 2: Lifeworld and system: A critique of functionalist reason. Boston, MA: Bacon Press (1987)

[34] Habermas, J.: Justification and application: Remarks on discourse ethics (1993)

[35] Haque, M.R., Saxena, D., Weathington, K., Chudzik, J., Guha, S.: Are we asking the right questions?: Designing for community stakeholders' interactions with ai in policing. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, Association for Computing Machinery, New York, NY, USA (2024)

[36] Harman, G.H.: The inference to the best explanation. The philosophical review **74**(1), 88–95 (1965)

[37] Henin, C., Le Métayer, D.: A framework to contest and justify algorithmic decisions. AI and Ethics **1**(4), 463–476 (2021)

[38] Henin, C., Métayer, D.L.: Beyond explainability: Justifiability and contestability of algorithmic decision systems. AI and Society **37**(4), 1397–1410 (2022). https://doi.org/10.1007/s00146-021-01251-8

[39] Hilton, D.J.: Conversational processes and causal explanation. Psychological Bulletin **107**(1), 65 (1990)

[40] Hohman, F., Head, A., Caruana, R., DeLine, R., Drucker, S.M.: Gamut: A design probe to understand how data scientists understand machine learning models. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–13 (2019)

[41] Hurlin, C., Pérignon, C., Saurin, S.: The fairness of credit scoring models. arXiv preprint arXiv:2205.10200 (2022)

[42] Juanico, R., Sarles, N.: Assemblée nationale, Rapport d'information sur l'évaluation de l'accès à l'enseignement supérieur, vol. 3232. Assemblée nationale (2020)

[43] Kasy, M., Abebe, R.: Fairness, equality, and power in algorithmic decision-making. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2021)

[44] Korikov, A., Shleyfman, A., Beck, C.: Counterfactual explanations for optimization-based decisions in the context of the gdpr. In: ICAPS 2021 workshop on explainable AI planning (2021)

[45] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., Wong, W.K.: Too much, too little, or just right? Ways explanations impact end users' mental models. In: 2013 IEEE Symposium on visual languages and human centric computing. pp. 3–10. IEEE (2013)

[46] Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: A joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1675–1684 (2016)

[47] Lanzetti, N., Dörfler, F., Pagan, N.: The impact of recommendation systems on opinion dynamics: Microscopic versus macroscopic effects. In: 2023 62nd IEEE Conference on Decision and Control (CDC). pp. 4824–4829. IEEE (2023)

[48] Leake, D.B.: Abduction, experience, and goals: A model of everyday abductive explanation. Journal of Experimental & Theoretical Artificial Intelligence **7**(4), 407–428 (1995)

[49] Lerouge, M., Gicquel, C., Mousseau, V., Ouerdane, W.: Modeling and generating user-centered contrastive explanations for the workforce scheduling and routing problem. International Transactions in Operational Research **33**(3), 1525–1558 (2026)

[50] Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)

[51] Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., Hardt, M.: Delayed impact of fair machine learning. In: International Conference on Machine Learning. pp. 3150–3158. PMLR (2018)

[52] Lowe, E.J.: Free agency, causation and action explanation. In: Sandis, C. (ed.) New Essays on the Explanation of Action, pp. 338–355. Palgrave Macmillan UK, London (2009)

[53] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[54] Meinard, Y., Tsoukiàs, A.: On the rationality of decision aiding processes. European Journal of Operational Research **273**(3), 1074–1084 (2019)

[55] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)

[56] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. arXiv preprint arXiv:1811.07867 (2018)

[57] Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM Transactions on Interactive Intelligent Systems (TiiS) **11**(3-4), 1–45 (2021)

[58] Mok, L., Nanda, S., Anderson, A.: People perceive algorithmic assessments as less fair and trustworthy than identical human assessments. Proceedings of the ACM on Human-Computer Interaction **7**(CSCW2), 1–26 (2023)

[59] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys **55**(13s), 1–42 (2023)

[60] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)

[61] Parikh, R.B., Teeple, S., Navathe, A.S.: Addressing bias in artificial intelligence in health care. Jama **322**(24), 2377–2378 (2019)

[62] Pearl, J.: Causality. Cambridge university press (2009)

[63] Perelman, C., Olbrechts-Tyteca, L.: Traité de l'argumentation, vol. 1. Presses universitaires de France (1958)

[64] Perelman, C.: Jugements de valeur, justification et argumentation. Revue Internationale de Philosophie **15**, 327–335 (1961)

[65] Rawls, J.: The idea of public reason revisited. The university of Chicago law review **64**(3), 765–807 (1997)

[66] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

[67] Ricœur, P.: La mémoire, l'histoire, l'oubli. L'Ordre philosophique, Le Seuil, Paris (2014)

[68] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)

[69] Saxena, D., Moon, S.Y., Shehata, D., Guha, S.: Unpacking invisible work practices, constraints, and latent power relationships in child welfare through casenote analysis. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–22 (2022)

[70] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

[71] Stevenson, M.: Assessing risk assessment in action. Minn. L. Rev. **103**, 303 (2018)

[72] von Wright, G.: Explanation and Understanding. Routledge, London (1971)

[73] Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International data privacy law **7**(2), 76–99 (2017)

[74] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech. **31**, 841 (2017)

[75] Waldman, A.E.: Power, process, and automated decision-making. Fordham L. Rev. **88**, 613 (2019)

[76] Walton, D.: Examination dialogue: An argumentation framework for critically questioning an expert opinion. Journal of pragmatics **38**(5), 745–777 (2006)

[77] Walton, D.: Dialogical models of explanation. ExaCt **2007**, 1–9 (2007)

[78] Walton, D.: A dialogue system specification for explanation. Synthese **182**(3), 349–374 (2011)

[79] Wick, M.R., Thompson, W.B.: Reconstructive expert system explanation. Artificial Intelligence **54**(1-2), 33–70 (1992)