LUISS

Data Science and Management

Course of Data Privacy and Security

From Black Boxes to Explainable Matchmaking: Transparency, Privacy, and Governance in Enterprise AI

Supervisor:

Paolo Spagnoletti

Co-supervisor:

Fabio Angeletti

Candidate:

Benedetta Sabatino

781701

Academic Year 2024/2025

Contents

1 Explainable Artificial Intelligence: Origins, Evolution and Interpretability Techniques 1

1.1 Introduction 1

1.2 Origins and Motivation 1

1.3 Black Box Problem 3

1.4 Evolution of XAI Techniques 4

1.5 Interpretability Techniques: Approaches and Debates 7

1.5.1 Types of Explainability 7

1.5.2 Limitations and Open Debates 9

1.5.3 Evaluating and Contextualizing Explanations 10

1.6 Explainable AI in Practice: Case Studies in Healthcare, Finance, HR, and Cybersecurity 11

1.6.1 Healthcare: Transparent AI for Clinical Decision Support 12

1.6.2 Finance: Explainability for Risk Management and Compliance . . 15

1.6.3 Human Resources: Explanations to Enhance Fairness and Feedback 18

1.6.4 Cybersecurity: Interpretable Insights for Threat Detection 22

1.6.5 Conclusion 27

2 Methodologies for B2B Matchmaking in Digital Business Ecosystems 29

2.1 Introduction to B2B Matchmaking 29

2.1.1 Definition and Context 29

2.1.2 Historical Evolution 30

2.1.3 Importance in Business Service Contexts 31

2.1.4 Overview of Methodological Approaches 31

2.1.5 Scope and Objectives 32

2.1.6 Methodology 1: Weighted Scoring Matrix 33

2.1.7 Methodology 2: Strategic Compatibility Framework 37

2.1.8 Methodology 3: Multi-Criteria Analysis (MCA) 42

2.1.9 Methodology 4: Profiling and Gap Analysis 47

2.1.10 Methodology 5: KPI-based Decision Matrix 53
2.2 Comparative Analysis 59
2.2.1 Comparative Framework 59
2.2.2 Detailed Comparison 59
2.2.3 Synthesis Matrix 62
2.2.4 A context-dependent framework for Methodology Selection 62
2.2.5 Hybrid Approaches 63
2.3 Conclusions 64
2.3.1 Synthesis of Key Insights 64
2.3.2 Implications for Practice 65
2.3.3 Future Directions and Transition to Application 66

3 Methodology **67**
3.1 Introduction to the Problem and Methodological Approach 67
3.2 The Dataset 68
3.2.1 Generation of the Synthetic Dataset 68
3.2.2 Exploratory Data Analysis: Understanding the Simulated Ecosystem 69
3.2.3 Data Preprocessing and Splitting: Translating Business Concepts into a Machine-Readable Form 71
3.3 Model Architecture and Training: Engineering a System for Semantic Compatibility 72
3.4 Evaluation Methodologies: Measuring Effectiveness and Ensuring Transparency 74
3.4.1 Measuring Effectiveness 74
3.4.2 Explainability Framework 75

4 Results and Discussion **79**
4.1 Validation in the Cyber 4.0 Context 79
4.2 Quantitative Results and Model Performance 82
4.3 Explainability and Transparency Analysis (XAI) 83
4.3.1 Introduction to XAI Methods 83
4.3.2 LIME: Output and Interpretation 83
4.3.3 LIME Results and Analysis 83

4.3.4 Integrated Gradients (IG): Output and Interpretation 84
4.3.5 IG Results and Interpretation 84
4.3.6 Conclusion 84
4.4 Limitations of the Study 85
4.4.1 Hardware Constraints and Experimental Scope 85
4.4.2 Data Strategy and Generalizability Risks 85
4.4.3 Scope and Nuances of Explainability 85
4.4.4 Expert Validation and Key Findings 86
4.4.5 Synthesis of the Contribution 86

5 Conclusions 87

5.1 Overview 87
5.2 Answer to the Research Question 87
5.3 Contributions and Implications 89
5.4 Future Vision 90
5.4.1 A Role-Based, Context-Aware System 90
5.4.2 Dual-Purpose Strategic Intelligence 90
5.4.3 Adaptive Explainability and Generative AI 90
5.4.4 From Explanations to Patterns 91
5.4.5 A Simple Post-Match Feedback Loop 91
5.4.6 Design Variants Where Appropriate 91
5.4.7 The End Goal: A Dynamic Strategic Partner 91
5.4.8 Closing Reflection 92

List of Tables and Figures

List of Tables

2.1 B2B Matchmaking – Methodology Summary 58

4.1 Test performance (held-out test) 82

4.2 Example LIME contributions for a "Match" instance 83

4.3 Example IG attributions for the same instance 84

List of Figures

3.1 Distribution of beneficiary sectors in the training set, showing a significant concentration in Manufacturing. 70

3.2 Distribution of key technological needs among beneficiaries, highlighting the prevalence of Big Data. 70

3.3 Distribution of technical competencies across the supplier population. . . 71

1

Explainable Artificial Intelligence: Origins, Evolution and Interpretability Techniques

1.1 Introduction

Artificial intelligence (AI) systems are increasingly being deployed in high-stakes domains such as healthcare, finance, and autonomous driving, where understanding how decisions are made is not just beneficial but essential. However, many modern AI approaches—particularly those based on machine learning and deep neural networks—function as opaque "black boxes," generating outputs without offering accessible or intuitive reasoning. This lack of transparency can significantly erode user trust and limit the integration of AI into mission-critical contexts. In response, the field of Explainable Artificial Intelligence (XAI) has emerged, aiming to develop models and tools that provide human-understandable explanations for AI decisions. The motivations behind XAI are both practical and normative: from facilitating model debugging and enhancing user confidence, to addressing ethical and legal imperatives such as fairness, accountability, and compliance with transparency-driven regulations.Overall, XAI plays a key role in bridging the gap between complex model behavior and the interpretability required for responsible and informed deployment.

1.2 Origins and Motivation

The need for interpretability in AI is not new. In the 1970s and 1980s, early expert systems, programs that used human knowledge encoded as rules, were already designed to explain their decisions. A well-known example is the MYCIN medical expert system, which could explain the reasoning behind its diagnoses to doctors[1]. These knowledge-based systems were inherently interpretable as they followed human-readable logical rules, and their explanations were “usually human-understandable” to users.

In the 1990s, however, AI research shifted toward machine learning and statistical models. This change brought significant improvements in performance, but it also introduced a problem: models like neural networks and ensembles became harder to interpret. Their internal logic was no longer based on explicit rules, making it difficult for humans to follow how decisions were made. As a result, the focus on built-in explainability declined. Interest in explainability re-emerged in the mid-2010s. The success of deep learning models, often described as “black boxes,” raised concerns about transparency. These models were accurate but difficult to understand, especially in high-stakes settings.

Regulatory and policy developments have further motivated XAI. In particular, the European Union’s General Data Protection Regulation (GDPR) has been widely interpreted to grant individuals the ’right to an explanation’ for algorithms’ decisions. Specifically, GDPR Article 22 and Recital 71 imply that when individuals are subject to automated decisions, they have the right to obtain “meaningful information about the logic involved”. This legal pressure in Europe pushed organizations to seek ways for AI systems to provide explanations to users or auditors.

In the United States, defense agencies also recognized the strategic importance of AI transparency. In 2016, the Defense Advanced Research Projects Agency (DARPA) launched the Explainable AI program. Its goal was to support the development of models that remain accurate but are easier for people to understand, trust, and manage. The program made it clear that explainability is essential if AI is to be used reliably in critical applications such as military operations[2].

The origins of XAI trace back to the explainability of early expert systems and the subsequent loss of transparency in statistical learning. The modern resurgence of XAI is fueled by the need for trust and accountability in complex AI systems and is reinforced by legal rights and government programs demanding transparency. Explainability is now seen as a key requirement for “responsible AI,” alongside principles like fairness and

[1] Bruce G Buchanan and Edward H Shortliffe. *Rule based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence).* Addison-Wesley Longman Publishing Co., Inc., 1984.

[2] David Gunning and David Aha. “DARPA’s explainable artificial intelligence (XAI) program”. In: *AI magazine* 40.2 (2019), pp. 44–58.

safety.[3]

1.3 Black Box Problem

The Black Box Problem in artificial intelligence (AI) refers to the opacity of certain advanced AI models whose internal workings and decision-making processes are not transparent or interpretable to humans. This challenge is particularly prevalent in models such as deep neural networks (DNNs) and complex ensemble methods; due to their extensive number of parameters, nonlinear structures, and intricate data interactions, these models can achieve remarkable predictive accuracy but at the expense of interpretability[4].

At the core of the Black Box Problem is the trade-off between predictive performance and model comprehensibility. Highly accurate models often depend on complex feature interactions and non-linear transformations, which make it difficult, if not impossible, for human analysts to intuitively grasp why a particular decision was made[5]. This lack of interpretability poses significant challenges across multiple dimensions: trust, accountability, fairness, and regulatory compliance. Users and stakeholders are typically hesitant to adopt or trust AI systems whose rationale behind critical decisions is unclear.[6]

Accountability becomes particularly problematic with black-box models in regulated industries where organizations must justify their decisions clearly and transparently. Furthermore, fairness concerns arise when opaque AI models may inadvertently learn biases embedded in training data, leading to systematically biased decisions without transparent ways to detect or mitigate these issues.

Explainable Artificial Intelligence (XAI) directly addresses the Black Box Problem by developing methods aimed at making model decisions understandable to humans. XAI approaches include both post-hoc explanations, which clarify model predictions after training, or use models that are transparent by design, like decision trees or rule-based systems[7].

[3] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". in: *Information fusion* 58 (2020), pp. 82–115.

[4] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[5] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[6] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[7] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable

While XAI methods have significantly advanced interpretability, debates persist regarding their sufficiency. Critics argue that post-hoc explanations may offer only approximate or misleading insights into model behavior, advocating instead for intrinsically interpretable models when high stakes are involved [6]. Therefore, addressing the Black Box Problem effectively demands careful balance between model complexity, interpretability, and practical constraints in the specific application domain.

1.4 Evolution of XAI Techniques

Over the past decade, XAI has evolved from a niche research interest into a broad, interdisciplinary field. Early efforts in the 2010s primarily focused on making sense of specific black-box models. Pioneering work by Ribeiro et al. (2016)[8] introduced Local Interpretable Model-Agnostic Explanations (LIME), which offered a general approach to explain any classifier's individual predictions. LIME works by perturbing an input and training a simple surrogate model (like a linear model) around the locality of that input to approximate the complex model's behavior; the surrogate's weights then serve as an explanation for which features are most influential in that particular prediction.

Around the same time, Lundberg and Lee (2017)[9] developed SHAP (SHapley Additive Explanations) as a unifying framework for feature importance based on game theory. SHAP assigns each feature a "credit" value for a given outcome, satisfying theoretical properties from Shapley values to ensure a fair allocation of importance [9]

These two methods—LIME and SHAP—became milestone contributions in XAI, widely cited and applied due to their model-agnostic nature (they can explain any type of model) and their intuitive output (highlighting feature contributions).

Another significant development in this period was the rise of counterfactual explanations. Rather than highlighting what features influenced a decision, a counterfactual explanation tells how a decision could have been different. Wachter, Mittelstadt, and Floridi (2017)[10] proposed counterfactual reasoning as a means to give end-users action-

artificial intelligence (XAI)". in: *IEEE access* 6 (2018), pp. 52138–52160.

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" 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.* 2016, pp. 1135–1144.

[9] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[10] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without open-

able recourse without revealing the model's inner workings, aligning with the spirit of GDPR's transparency requirements. For example, a counterfactual explanation to a loan applicant might be: "Had your income been $5,000 higher, the loan would be approved" – pointing out a minimal change to achieve a different result. This approach gained traction for being user-centric and legally accepted: it provides a reason in terms of a possible action by the individual, rather than exposing the algorithm's code or weights[11].

In parallel, the deep learning boom spurred techniques to interpret neural networks, particularly in computer vision. Saliency maps or heatmaps were introduced as a way to visualize which parts of an input (e.g. regions of an image) most affect a model's output. Simonyan et al. (2014)[12] demonstrated that by taking the gradient of the output with respect to input pixels, one can produce a crude "attention map" highlighting important image regions for a classification.

Later methods like Grad-CAM[13] improved on this by using internal layer activations to create more focused localization maps for a given predicted class, making it clearer why a convolutional network chose a particular label. Grad-CAM (Gradient-weighted Class Activation Mapping) works by computing the gradient of the target class with respect to the feature maps of the last convolutional layer, weighting these maps by the average gradient, and combining them to produce a low-resolution heatmap that highlights the important regions in the input image. This visualization technique enables practitioners to identify which parts of the input most influenced the model's decision. Such interpretability tools allowed researchers and practitioners to peer into the "attention" of deep models, sometimes revealing that models were picking up on intuitive features (e.g., focusing on a tumor region in a medical image), or in other cases, revealing misleading signals (e.g., focusing on irrelevant background patterns), which could prompt corrective action.

By the late 2010s, XAI had matured with an increasing number of reviews and taxonomies that attempted to organize the developing landscape of methods. Researchers

ing the black box: Automated decisions and the GDPR". in: *Harv. JL & Tech.* 31 (2017), p. 841.

[11] Roberto Confalonieri et al. "A historical perspective of explainable artificial intelligence". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.1 (2021), e1391.

[12] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[13] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision.* 2017, pp. 618–626.

distinguished between global explanations (interpreting an entire model's logic or behavior) and local explanations (interpreting individual predictions), and among explanations aimed at experts (model developers troubleshooting a system) versus lay users (end-users affected by a decision). There was also increasing interplay between XAI and other ethical AI concerns. One notable trend was the integration of explainability with efforts to ensure algorithmic fairness and to mitigate bias. It was hoped that providing explanations for model decisions would make it easier to detect and correct biased or unfair outcomes[14]. For instance, an explanation might reveal that a credit scoring model heavily relies on a customer's zip code – which could be a proxy for race – thus flagging potential discrimination. However, recent critical studies have cautioned against viewing XAI as a silver bullet for fairness. While literature is "highly optimistic" about XAI's fairness benefits for various stakeholders, the reality is more nuanced. Explanations can certainly help stakeholders understand and scrutinize AI decisions, but they do not automatically resolve underlying biases, and in some cases they may even mislead or create false assurances of fairness. As highlighted by Deck et al. (2024)[15], "Explainable AI should not be conceived as an ethical panacea but as one of many tools for addressing the multifaceted challenge of algorithmic fairness", underlining the importance of contextualizing XAI within a broader ecosystem of fairness-enhancing strategies.

Since 2020, several new frontiers have emerged in XAI. One is the challenge of explaining large language models (LLMs) and other advanced generative AI systems. Models like GPT-3 and GPT-4, with billions of parameters trained on internet-scale data, are extraordinarily complex. Researchers have begun adapting XAI methods or inventing new ones to interpret these models' inner workings and outputs. For example, recent studies introduce taxonomies of techniques for explaining the decisions and internal representations of transformer-based LLMs (e.g. by analyzing attention patterns or hidden unit activations).

Another emerging area is explainable generative AI (GenXAI). Generative models (like GANs and diffusion models for images) can produce highly realistic outputs, but

[14] Navita Goyal et al. "The impact of explanations on fairness in human-ai decision-making: Protected vs proxy features". In: *Proceedings of the 29th International Conference on Intelligent User Interfaces.* 2024, pp. 155–180.

[15] Luca Deck et al. "A Critical Survey on Fairness Benefits of Explainable AI". in: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24).* Accessed: 2025-04-06. ACM, 2024. URL: https://facctconference.org/static/papers24/facct24-105.pdf.

understanding why a particular output was generated or how to control such models is challenging. New research in GenXAI seeks to provide explanations for generative processes – for instance, linking certain latent variables to aspects of the output image, or generating counterfactual images to show what changes in input lead to different outputs[16]. Additionally, there is growing recognition of the human-centered aspects of explainability: social scientists and Human-Computer Interaction researchers are studying how people perceive and use AI explanations. For XAI to fulfill its promise, explanations must not only be technically accurate and faithful to the model, but also intelligible and useful to the target audience. This has led to user studies on explanation interfaces, research on how explanations affect trust and decision-making, and considerations of how to tailor explanations to different users (a doctor might need a different type of explanation than a patient, for example). The evolution of XAI thus spans technical innovation, practical integration into AI workflows, and interdisciplinary research into the effectiveness and ethics of explanations.

1.5 Interpretability Techniques: Approaches and Debates

1.5.1 Types of Explainability

XAI techniques can be broadly classified along several dimensions. One key distinction is between intrinsic interpretability and post-hoc explainability. Intrinsically interpretable models are those that are transparent by design – their structure or parameters can be directly inspected and understood without additional tools. Classic examples include decision trees (which present human-readable decision paths), rule-based systems, linear or logistic regression models (where each feature weight is explicit), and generalized additive models (which extend linear models with intuitive non-linear contributions for each feature). Because these models themselves serve as their own explanation, they have the advantage of high fidelity (the explanation is identical to the model's reasoning) and often straightforward simulatability (a human can manually step through the model's logic)[6]. However, they may sacrifice predictive power on complex tasks; there is often a

[16] Johannes Schneider. "Explainable generative ai (genxai): A survey, conceptualization, and research agenda". In: *Artificial Intelligence Review* 57.11 (2024), p. 289.

trade-off where simpler models are more interpretable but less accurate than black-box models like deep neural networks [9]. Modern research is trying to soften this trade-off by developing more expressive yet interpretable models – for example, optimized decision sets or rule lists that remain small and comprehensible, or hybrid models that combine transparent components with black-box components in a controlled way.

In contrast, post-hoc explanation methods are applied after a model has been trained, to extract explanations for its predictions without altering the model's internal workings[6]. These methods treat the original model as given (often an inscrutable complex model) and produce an approximation or annotation that is easier to understand. Post-hoc techniques fall into two main types. Model-agnostic methods treat the model as a black box and can be applied to any system. Model-specific methods use internal information to tailor the explanation. LIME and SHAP, discussed earlier, are examples of model-agnostic methods: they require only the ability to query the model's output for different inputs, not how the model internally arrives at its output[9]. They typically provide local explanations – e.g. explaining one individual prediction at a time by estimating feature influence in the vicinity of that data point. Other model-agnostic tools include partial dependence plots and individual conditional expectation (ICE) plots, which show how changing one feature (or a pair of features) while holding others fixed affects the model's prediction, thus offering a more global view of a model's behavior with respect to that feature. There are also example-based explanations: for instance, retrieving prototypes (representative cases from the training set that the model finds similar to the input) or influential examples (training examples that strongly influence the model's prediction) can help explain decisions by analogy. Counterfactual explanations, while model-agnostic in formulation, often require solving an optimization to find a close alternative input that flips the model's outcome, which can be done either by repeatedly querying the model or by using gradient information if available.

Model-specific explanation techniques leverage internal access to model operations to generate detailed insights. In neural networks, a prominent class of methods is based on gradient-based feature attribution. Beyond basic gradient-based explanations, more robust approaches like Integrated Gradients[17] accumulate gradients along a defined path from a baseline input to the actual input, resulting in more stable and meaningful attri-

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning.* PMLR. 2017, pp. 3319–3328.

butions for individual features. Similarly, methods such as DeepLIFT[18] and Layer-wise Relevance Propagation[19] backpropagate the contributions from the output to input layers to quantify the influence of each input feature precisely. Another explanatory strategy involves constructing a surrogate model to mimic the behavior of an opaque model. For instance, decision trees or sets of rules can be trained on the inputs and outputs of complex neural networks, providing an approximate yet interpretable representation of the model's overall decision-making process. For ensemble-based models, such as random forests or gradient boosting machines, feature importance scores can be computed by measuring the extent to which each feature reduces prediction error across individual trees. Additionally, SHAP values offer an efficient method specifically tailored for tree-based models, quantifying feature contributions in a consistent manner. Certain techniques are tailored explicitly to specific model architectures. For example, attention-weight visualizations in transformer models clarify model behavior by highlighting which input segments receive the most focus. In the realm of reinforcement learning, explanation methods often extract features critical to reward predictions or simulate alternative action paths to explain the learned policy decisions clearly.

1.5.2 Limitations and Open Debates

Each approach to interpretability has its use cases and limitations. Intrinsic interpretable models are favored in applications where transparency and simplicity are crucial and the problem complexity is moderate – for instance, a healthcare scorecard that doctors use might deliberately be a sparse linear model so that the rationale for a score is obvious and can be trusted. Post-hoc explanations are indispensable when one must deploy a complex model (for its accuracy or scalability) but still needs to justify or audit its decisions – for example, explaining the predictions of a proprietary credit-risk model to satisfy customer inquiries or regulatory oversight. However, post-hoc methods come with the caveat that their explanations are approximations of the model's reasoning. A common critique is that they may not be faithful to what the model truly computes, especially if the model has complex non-linear interactions that a simple explanation

[18] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences". In: *International conference on machine learning.* PMlR. 2017, pp. 3145–3153.

[19] Sebastian Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

cannot capture[6]. Ribeiro et al. [8] acknowledged this, noting that while LIME explanations are locally faithful, they are not guaranteed to work globally for the model's decision boundary. Research has also shown that methods like LIME and SHAP can be unstable: small changes in the input or random initialization can yield different explanations, raising concerns about their reliability. In fact, adversarial attacks on explanation methods have been demonstrated – subtly perturbing a model or its input to produce misleading explanations that hide the model's true reasoning or falsely highlight irrelevant factors[20]. These findings underscore that explanations should ideally be accompanied by some measure of confidence or robustness.

Beyond technical validity, discussions continue regarding the effectiveness and ethical implications of Explainable AI (XAI) techniques. A central debate is encapsulated in Cynthia Rudin's argument [6] that for high-stakes decision-making, we should "stop explaining black-box models" and instead adopt interpretable models from the outset. Rudin argues that post-hoc explanations of complex models often fail to reliably reflect the model's actual decision-making process, potentially providing misleading or speculative insights. Consequently, she recommends inherently interpretable models—such as sparse scoring systems or rule-based methods—in domains like medical diagnosis or criminal justice risk assessment, where transparency about each input's contribution is paramount. This stance has stimulated significant academic discourse. Some researchers agree that interpretable models can often match or approximate the performance of black-box models while offering superior transparency, safety, and accountability. Conversely, other scholars highlight contexts where black-box models offer substantial performance advantages, asserting that despite their limitations, XAI approaches represent a crucial intermediary, balancing performance and interpretability.

1.5.3 Evaluating and Contextualizing Explanations

Another ongoing discussion addresses the evaluation criteria for effective explanations. Stakeholders often emphasize different qualities: data scientists might prioritize fidelity and completeness—ensuring explanations accurately represent the model's logic—while

[20] Ferhat Sarikaya. *The Quest for Explainability in Artificial Intelligence: Challenges, Progress, and Future Directions*. Accessed: 2025-04-04. 2024. URL: `https://medium.com/@ferhatsarikaya/the-quest-for-explainability-in-artificial-intelligence-challenges-progress-and-future-3e36626d58ae`.

end-users typically prefer clarity and actionable insights, guiding practical decisions and fostering trust [11]. There remains no universally accepted set of metrics for assessing explanation quality. Researchers have proposed several criteria, including accuracy (how correctly the explanation predicts model outputs), stability (consistency of explanations for similar inputs), comprehensibility (ease of understanding by non-experts), and practical usefulness (ability to support improved decision-making and appropriate trust). Evaluations frequently involve human-subject studies, highlighting the inherently interdisciplinary nature of XAI research.

Finally, the integration of XAI with broader societal and ethical considerations continues to evolve. As noted, simply adding an explanation does not guarantee an AI system is fair or trustworthy, but it can be an important enabler for those properties. Explanations might help uncover biased reasoning (e.g. revealing that a hiring algorithm is implicitly using gendered features), but there is also a risk of "fairwashing" or "explanation washing", where superficial explanations give a false impression of fairness or diligence. Policymakers are beginning to outline standards for meaningful explanations – for instance, what counts as a valid explanation to fulfill legal requirements like the GDPR's call for "logic" disclosure. Audience and context are crucial: the appropriate form of explanation may differ for an AI developer debugging a model versus a consumer questioning an adverse decision. Moving forward, many suggest XAI should be part of a larger framework of AI governance, complementing techniques for bias mitigation, privacy protection, and accountability. Explainability, in this view, is not the goal. It's a tool to support better oversight and build trust in how AI is used.

1.6 Explainable AI in Practice: Case Studies in Healthcare, Finance, HR, and Cybersecurity

The implementation of Explainable Artificial Intelligence (XAI) has transitioned from theoretical exploration to practical application across various sectors. Building upon the foundational concepts and methodologies of XAI previously discussed, this section investigates its application in healthcare, finance, human resources, and cybersecurity, with a focus on case studies of implemented systems. The analysis includes an examination of the XAI techniques employed, such as saliency maps, SHAP values, and counterfactual explanations, the rationale behind their selection, and their integration into existing workflows. Additionally, the impact of these techniques on trust, accuracy, fairness, and compliance is assessed. This case-driven analysis elucidates how the potential of XAI is realized in operational contexts and the lessons learned for enhancing transparency and accountability in AI decision-making within high-stakes domains.

1.6.1 Healthcare: Transparent AI for Clinical Decision Support

Healthcare has seen some of the most safety-critical deployments of XAI. In hospitals and clinics, AI aids in tasks like medical image interpretation and diagnosis, but clinicians demand to "see inside" these black boxes before trusting them. A notable implementation is in medical imaging diagnostics: many AI systems now provide saliency map overlays on images to highlight regions that influenced a prediction. For example, an AI model detecting pneumonia on chest X-rays might output a heatmap on the image, marking the suspicious lung opacity that led to the diagnosis[21]. These visual explanations help radiologists verify that the AI is focusing on medically relevant features (e.g. an infiltrate in the lung) rather than noise. Indeed, explainability has been shown to improve physician trust and acceptance of AI: having an "AI assistant that not only makes recommendations but also explains its reasoning in clear, medical terms" bridges the gap between complex models and clinical intuition [21].

One real-world application of Explainable AI (XAI) is at Moorfields Eye Hospital in London, where researchers developed a deep learning system for detecting retinal diseases. They implemented an inherently interpretable two-stage approach: the AI first generates anatomical segmentations of optical coherence tomography (OCT) scans, identifying clinically relevant features such as lesions and abnormal fluid accumulations. Subsequently, the system provides a referral recommendation based explicitly on this segmented anatomical data[22]. This design choice was intentional, ensuring clinicians always have direct visibility into the reasoning behind the AI's decisions. The intermediate segmentation clearly illustrates clinical findings, enabling ophthalmologists to verify the

[21] SmythOS. *Top Use Cases of Explainable AI: Real-World Applications for Transparency and Trust.* Accessed: 2025-04-04. 2025. URL: `https://smythos.com/ai-agents/agent-architectures/explainable-ai-use-cases/`.

[22] Jason Yim et al. *Using AI to predict retinal disease progression.* Accessed: 2025-04-04. Google DeepMind. 2020. URL: `https://deepmind.google/discover/blog/using-ai-to-predict-retinal-disease-progression/`.

exact anatomical features—such as drusen deposits or retinal thickness changes—that underpin the referral decision. Clinicians reported increased confidence in these AI-generated recommendations, as the transparent anatomical explanations allowed them to validate and contextualize the AI's assessments effectively. Moreover, the risk scores provided by the system were consistent with observed anatomical changes over time, offering clinicians deeper insights into disease progression. This integration of explainability into clinical practice not only satisfied regulatory expectations for meaningful explanations but also significantly enhanced clinician trust and maintained high diagnostic accuracy.

Another hospital deployment of XAI is in intensive care units, where predictive models (for patient deterioration, sepsis risk, etc.) are used by clinicians. For instance, a sepsis early-warning system might use feature attribution methods (like SHAP values) to explain its risk score for a given patient. If an AI alert warns that a patient is high-risk for sepsis, the system can display a ranked list of contributing factors – e.g. "very low blood pressure, high heart rate, and rising lactate were the top contributors to the sepsis prediction" – with each factor's influence quantified[23]. Such explanations were chosen to integrate into existing electronic health record dashboards, so that doctors and nurses can immediately see why the AI is sounding an alarm. In practice, this has improved user trust and adherence to AI alerts: instead of a mysterious score, clinicians get a narrative like "the model thinks this patient may be developing sepsis because of X, Y, and Z," which they can confirm against their own examination. Hospital case studies have reported that when explanations align with clinical reasoning, staff are more likely to act on AI recommendations, potentially improving outcomes[23].

Despite these successes, healthcare XAI deployments have also revealed limitations and debates. Saliency maps, while popular, do not always guarantee a faithful explanation. An illuminating example emerged during the COVID-19 pandemic: a flurry of AI models were developed to detect COVID-19 pneumonia from chest X-rays, and many boasted high accuracy. However, researchers at MIT and partners applied explainability techniques to these models and discovered troubling "shortcuts" – the AI was often look-

[23] Nvidia and Centre for Data Ethics and Innovation (CDEI). *Explainable AI for Credit Risk Management: Applying Accelerated Computing to Enable Explainability at Scale for AI-powered Credit Risk Management Using Shapley Values and SHAP.* Accessed: 2025-03-04. Gov.uk. June 2023. URL: `https://www.gov.uk/ai-assurance-techniques/nvidia-explainable-ai-for-credit-risk-management-applying-accelerated-computing-to-enable-explainability-at-scale-for-ai-powered-credit-risk-management-using-shapley-values-and-shap`.

ing at irrelevant image features correlated with COVID status (such as hospital logos or patient position markers) rather than true pathology[24].In one case, the saliency maps for a COVID-positive X-ray lit up a corner of the image where a hospital's radiography label appeared, indicating the model had learned to cheat based on where the image came from. These findings, enabled by XAI (Grad-CAM heatmaps in this case), alerted doctors and developers to the lack of genuine medical signal in some models. As a result, several hospitals refrained from deploying such models in practice, underscoring that explanations can be essential for model validation: they can detect when an AI's high accuracy is deceptive and not rooted in the right reasons[24]. More generally, recent studies have shown that many saliency methods have poor robustness – slight, irrelevant changes to an input can alter the heatmap even if the model's prediction doesn't change. In a 2023 Radiology study, seven saliency techniques were tested on a chest X-ray model and all failed at least one sanity check for reliability. The authors concluded that saliency maps can "misrepresent [the] true model prediction", cautioning clinicians not to take them at face value[25]. This has sparked debate: some argue that we need better, more rigorous XAI methods for medicine, while others suggest using inherently interpretable models (like sparse risk scores or causal models) for critical healthcare applications. There are discussions around attention mechanisms in medical AI (for example, attention-based models highlight important symptoms in clinical text) – are these attention weights truly offering an explanation, or just another ad hoc visualization? Researchers Jain and Wallace famously found that in NLP models, "attention is not explanation," since one can often alter the attention weights without changing a model's output[26]. Such insights urge caution: even as XAI improves transparency in healthcare, practitioners and regulators remain wary of over-reliance on imperfect explanations. The overall lesson is that XAI adds value in clinical AI deployments by fostering trust and accountability, but it must be applied with an understanding of its limitations. When done thoughtfully, as in the Moorfields example or ICU alerts, XAI can meaningfully align AI systems with the

[24] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. "AI for radiographic COVID-19 detection selects shortcuts over signal". In: *Nature Machine Intelligence* 3.7 (2021), pp. 610–619.

[25] Jiajin Zhang et al. "Revisiting the trustworthiness of saliency methods in radiology AI". in: *Radiology: Artificial Intelligence* 6.1 (2023), e220221.

[26] Sarah Wiegreffe and Yuval Pinter. "Attention is not not explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, pp. 11–20.

interpretative, evidence-based mindset of healthcare professionals.

1.6.2 Finance: Explainability for Risk Management and Compliance

The financial industry was quick to embrace XAI due to strict regulations on fairness and transparency in decisions like credit lending, trading, and fraud detection. In banking, an oft-cited success story is the integration of SHAP (Shapley Additive Explanations) into credit risk models. SHAP assigns each feature a contribution value for a given prediction, which is particularly useful for explaining why a loan application was approved or denied. For example, if an AI model declines a customer's loan, the bank can generate a SHAP explanation listing factors such as "High debt-to-income ratio (-0.25 to approval probability), Short credit history (-0.15), and Stable employment (+0.10)" – indicating how each factor nudged the decision [23]. Banks have selected SHAP because it provides consistent and local explanations grounded in game theory (treating the model's prediction as a payoff distributed among features). In practice, these SHAP-based explanations are delivered to various stakeholders: loan officers see them to understand model decisions, customers receive them as reason codes in adverse action letters, and regulators review them to ensure compliance with consumer protection laws [23].According to a 2020 case study reported by NVIDIA [23], a major European bank implemented a SHAP-supported credit scoring system, which significantly improved communication with regulators by enabling clear explanations for AI-driven credit risk assessments, fulfilling the "explainability profile" required by law. However, a challenge was that calculating SHAP values for complex models at scale can be computationally intense (originally taking hours or days for large portfolios). To overcome this, an industry-academia collaboration led by NVIDIA introduced GPU-accelerated SHAP for credit risk management. By optimizing SHAP computations on parallel hardware, they achieved near real-time explanations even for thousands of loans, "enabling financial institutions to generate the explainability profile of entire portfolios in minutes rather than days"[23]. This breakthrough made it feasible to integrate SHAP into live risk monitoring systems – for instance, a risk manager can now run an overnight batch that not only flags high-risk loans but also produces an explanation for each. The use of XAI here had a direct impact on trust and adoption: previously, some banks were hesitant to deploy more accurate (but opaque) AI models for credit risk, fearing they could not justify decisions to auditors. With fast SHAP explanations, those same models became auditable and tractable, bridging the “explainability gap” in financial AI. Senior management and compliance teams could get global variable importance reports (which features generally drive defaults) as well as local explanations for specific cases, all traced and documented. This traceability is crucial – one bank combined the SHAP output with a documentation pipeline to maintain an “unambiguous record of the decisions made by the AI system,” including the key variables for each decision. Such integration of XAI satisfies both business needs (confidence in model behavior) and regulatory requirements (e.g. EU’s GDPR and banking regulators’ demands for algorithmic accountability), thereby driving AI adoption in finance.

Beyond credit scoring, fraud detection and anti-money laundering (AML) systems have incorporated explainability to improve their effectiveness. Fraud and AML models often produce alerts that human investigators must review, so feature attribution helps provide rationale for each alert. American Express, for example, handles over $1 trillion in transactions per year and uses an XAI-enhanced fraud detection system that explains why a transaction was flagged[21]. If an AI flags a credit card charge as fraudulent, the system might show a dashboard: “Suspicious because IP address = overseas (high risk), Amount is much larger than usual (moderate risk), and Card present = No (online transaction).” American Express analysts reported that these explanations help them pinpoint patterns and make quicker decisions on which alerts are true fraud[21].

This also aids in model tuning: if an explanation consistently points to a certain feature as top contributor, analysts investigate whether that feature is truly discriminative or causing too many false alarms[27]. In the AML domain, United Overseas Bank (UOB) in Singapore partnered with an AI startup to deploy a machine learning system for flagging money laundering transactions. A key design principle was that the system be “scalable and explainable”, allowing compliance officers to understand each flag[28]. They used a combination of SHAP and rule-based explanations to translate model outputs into understandable business terms for investigators and regulators. For instance, an alert might

[27] Transmit Security. *Solving AI’s Black-Box Problem with Explainable AI and SHAP Values.* https://transmitsecurity.com/blog/solving-ais-black-box-problem-with-explainable-ai-and-shap-values. Accessed: 2025-03-04. Apr. 2023.

[28] Deloitte. *The Case for Artificial Intelligence in Combating Money Laundering and Terrorist Financing.* Tech. rep. Accessed: 2025-03-04. Deloitte, June 2020. URL: https://www2.deloitte.com/content/dam/Deloitte/jp/Documents/financial-services/bk/en-the-case-for-artificial-intelligence-in-combating-money-laundering-and-terrorist-financing.pdf.

come with a note: "This transfer triggered because it's 5× the customer's usual amount and involves a high-risk jurisdiction." According to Deloitte's independent assessment of the UOB pilot, such explainability features "increased trust with end users and regulators", since decisions could be backed up with clear reasoning[28]. It also improved fairness and consistency: by explaining why certain customers were flagged for enhanced due diligence, the bank could ensure similar cases were handled equivalently and check that no prohibited attributes (e.g. race) were implicitly influencing the model.

An interesting effect of deploying XAI in finance is on the culture of model governance. Traditionally, financial institutions relied on simple, transparent scorecards for credit decisions (human-understandable but less accurate). The introduction of complex AI models necessitated new governance techniques – XAI tools became a core part of the process to validate and monitor these models. Banks now conduct "explainability performance testing" as part of model validation: for a sample of decisions, they inspect the XAI outputs and see if they align with domain knowledge [23]. If an explanation seems odd (e.g. a loan approved mainly due to an applicant's zip code), that prompts investigation into potential data biases or spurious correlations. This way, XAI is not just a post-hoc add-on but an integral assurance mechanism. Regulators have encouraged this – the UK's Financial Conduct Authority and others have issued guidance that AI in finance should have "appropriate transparency and explainability", proportional to the risk. In response, many banks have adopted a "human-in-the-loop" approach, where under certain conditions (such as borderline decisions or customer disputes) a human reviews the AI's explanation before finalizing the decision[23]. This hybrid approach, empowered by XAI, has shown improvements in trust (customers are more comfortable knowing a human reviewed an AI decision with an explanation) and accountability (the institution can justify decisions to an audit committee or in court if needed).

Despite the largely positive reception of XAI in finance, there are critical voices here as well. Some experts worry that post-hoc explanations like SHAP, while useful, might create an illusion of understanding. They point out that SHAP can explain a model's output in terms of input features, but it doesn't guarantee the model is reasoning in a causal or sensible way – much like how a complex derivative's risk can be summarized by numbers but still fail under unseen conditions.

Proponents of XAI in finance counter that some complex patterns (e.g. fraud rings, or subtle interactions predictive of default) truly require advanced models, and XAI is the compromise that allows those models to be deployed responsibly. Another concern is that explanations can be manipulated: researchers have shown it's possible to adversarially adjust a model so that the explanations hide biases or appear "fair" even when the model is unfair. This calls for careful validation of explanation fidelity in finance – for instance, stress-testing explanations by perturbing inputs and seeing if the explanations change accordingly (a practice some banks now follow, akin to sensitivity analysis). Overall, in finance the consensus is that XAI has become indispensable for aligning AI with the sector's longstanding demands for transparency, fairness, and control. It has enabled a new generation of AI systems that are more accurate yet still accountable, as evidenced by widespread adoption in credit scoring, portfolio management, and fraud analytics. The continued challenge will be ensuring these explanations remain reliable and that stakeholders don't become overconfident in AI just because it "explains itself" – a balance that ongoing research and regulation are actively trying to strike.

1.6.3 Human Resources: Explanations to Enhance Fairness and Feedback

In the human resources sector, AI tools are used for screening resumes, ranking job candidates, and even evaluating employee performance. These uses have raised sensitive issues of bias and fairness, making explainability crucial. A prominent application of XAI in HR is the generation of counterfactual explanations for hiring decisions. Unlike feature importance (which might say "Skill X and Y were most important in your score"), a counterfactual explanation provides a what-if scenario: it tells an applicant how the outcome would change if certain attributes were different. For example, imagine an AI-powered recruitment system that rejected a candidate for a data analyst role. A counterfactual explanation to the candidate might be: "If you had proficiency in Excel and Tableau, your application would be assessed as suitable for the Business Analyst position."[29]. This was demonstrated in a case study with a public employment service in Belgium, where an AI model scored job seeker–job posting matches. The XAI module (using a method called SEDC for generating counterfactuals) was able to output specific

[29] Raphael Mazzine et al. "Counterfactual explanations for employment services". In: *International workshop on Fair, Effective And Sustainable Talent management using data science.* 2021, pp. 1–7.

recommendations – effectively showing which skills the job seeker could add to become a good match. In one scenario, a candidate with a business administration background was not considered a fit for a data analyst job until the system suggested adding those software skills; once “Excel” and “Tableau” were hypothetically added to the resume, the model’s prediction changed from “not suited” to “suited”. HR departments selected this XAI approach to integrate into online career portals so that rejected candidates receive actionable feedback rather than a generic rejection. The impact has been positive on candidate experience: instead of feeling in the dark or assuming discrimination, candidates see a transparent (and constructive) reasoning. This transparency can also build trust in the AI – people are more likely to accept an algorithm’s decision if given a understandable reason and advice on how to improve. HR systems have also employed counterfactual explanations internally to audit and improve models. One famous incident was Amazon’s experimental hiring AI that inadvertently learned to penalize resumes containing the word “women” (as in “women’s chess club captain”)[30]. That model was never deployed due to bias, but it highlighted how easily AI can inherit past prejudices. To guard against this, organizations now use XAI techniques to probe their hiring algorithms for bias. In a retrospective analysis of Amazon’s case, researchers showed that a counterfactual explanation could have revealed the bias: “If you remove the words ‘women’s’ and ‘all-women university’ from the resume, the prediction changes from not selected to selected.”[29]. In other words, the model’s decision hinged on terms closely linked to gender – a clear red flag. Today, many recruitment AI vendors incorporate such bias detection tests. They will generate counterfactuals or contrastive explanations to check, for example, that changing the gender or ethnicity of a candidate (when not relevant to the job) does not affect the recommendation. This kind of XAI-driven auditing is increasingly mandated by regulations. The EU’s General Data Protection Regulation (GDPR) and recent laws in the U.S. (such as New York City’s bias audit requirement for hiring tools) push companies to explain and justify their automated hiring decisions, under threat of legal penalties. XAI helps companies demonstrate compliance by producing documentation on how their AI evaluates candidates and by ensuring any problematic correlations are caught early. One HR software provider, Workday, has gone as far as acquiring a startup (HiredScore) known for its explainable AI in talent screening. Workday stated

[30] **ilsif2023linkedin**).

that “HiredScore’s expertise in explainable AI” will be leveraged to make their recruiting recommendations more transparent and fair[31]. In practical terms, this means a hiring manager using Workday’s AI might see a list of top candidates where each candidate’s score comes with a brief explanation like “Match A – 90%: 5 years experience in required field, certification XYZ, past work at similar company.” The system might also flag why a candidate was not recommended, e.g. “Candidate B lacked certification XYZ (a key requirement)” – information that can be relayed to external recruiters or even to the candidate as constructive input (without revealing sensitive details or proprietary model information). The use of natural language explanations and simple grades (some tools use an A, B, C rating for candidate fit, accompanied by reasons) has been found effective for busy HR professionals who need quick insights rather than raw numbers[32]. It’s a design choice to ensure XAI integration doesn’t overwhelm users but instead streamlines decision-making.

Another area in HR is employee retention and performance management. Companies are using AI to predict employee attrition (who might quit) so that management can intervene. Explainability plays a role here too. If an AI flags an employee as “high flight risk,” HR wants to know why – is it low job engagement? skills mismatch? lack of recent promotion? One deployed system provided counterfactual reasons such as: “If the employee were given a 10% salary increase or moved to a different role utilizing skill X, the probability of attrition would drop significantly.” This kind of insight directly informs retention strategies (perhaps the solution is to offer a raise or a new project). A study by Mazzine et al. (2021)[29] demonstrated this by generating diverse counterfactual explanations for employees predicted to leave, suggesting different “what-if” scenarios to improve retention. These were presented as recommendations to HR (e.g. “Adding flexible working hours could change this person’s attrition risk from high to low”), which increased the actionability of the predictive model. Essentially, XAI turned a prediction into a set of possible solutions. Early adopters in large firms reported better trust in the model’s accuracy because the explanations often made sense and matched managers’ own intuitions about team members. Moreover, it helped to avoid blindly following AI:

[31] OutSail. *How Workday’s Acquisition of HiredScore is Transforming HR Tech.* Accessed: 2025-04-04. Nov. 2024. URL: `https://www.outsail.co/post/workdays-acquisition-of-hiredscore-reshaping-hr-technology`.

[32] Workday. *Responsible AI and Bias Mitigation.* Accessed: 2025-04-04. 2025. URL: `https://www.workday.com/en-us/legal/responsible-ai-and-bias-mitigation.html`.

if an explanation seems off (say it suggests something nonsensical), HR can question the model, check data quality or even decide to ignore a prediction – a safety check that pure black-box models wouldn't afford.

Despite these advantages, HR teams remain cautious due to the high stakes of fairness and ethics. XAI in HR is primarily seen as a tool to mitigate bias, but there's recognition that explanations alone don't guarantee fairness. Some critics argue that post-hoc explanations in hiring can give a false sense of security – a biased model might still produce "plausible" explanations. Therefore, many experts advocate combining XAI with rigorous bias audits and sometimes simpler models. For instance, if a complex AI is used for resume screening, one might also use an interpretable checklist model in parallel as a benchmark. If the black-box and the transparent model disagree wildly, the XAI explanations can help diagnose why. There's also the matter of candidate perception: while feedback is appreciated, not all candidates trust an algorithm's reasoning. If an explanation is poorly phrased or too generic ("you lacked a qualification"), it might not satisfy a rejected candidate and could even open the company up to disputes ("I do have that qualification, so your AI is wrong!"). Hence companies carefully design the content of explanations, often having legal and bias experts review them. The general trend, however, is toward greater transparency in HR AI. XAI deployments in recruitment have coincided with improved diversity outcomes in some cases – because when you see why the AI is picking certain candidates, you can notice if it's systematically favoring certain backgrounds and take corrective action (e.g. adjusting the model or input data). The explainable outputs also encourage a "feedback loop": recruiters can give feedback on whether an explanation was valid ("This candidate was indeed strong in X skill, good call") or not, which can be used to refine the AI system. Overall, XAI is helping transform HR AI from a mysterious filter into a more collaborative decision support tool. By providing user-centric explanations (to hiring managers, HR analysts, and candidates), these systems aim to ensure that algorithmic decisions in hiring and employment are justifiable, understandable, and improvable, aligning with both ethical practices and emerging legal standards.

1.6.4 Cybersecurity: Interpretable Insights for Threat Detection

Cybersecurity often involves rapidly analyzing data to detect threats like intrusions, malware, or fraudulent behavior. AI and machine learning models are increasingly deployed in security operations centers (SOCs) to sift through network logs, user activity, and transactions to flag anomalies. However, security analysts are notoriously skeptical of tools that raise frequent false alarms without explanation – an opaque model that raises an alert with no context is likely to be ignored. Thus, explainability has become a key feature in modern intrusion detection and cybersecurity AI systems.

A concrete example is in network intrusion detection systems (IDS). Consider an IDS that uses a trained model to identify potentially malicious network traffic among millions of connections. When it flags a certain network session as an attack, feature attribution can be used to explain why. For instance, the system might highlight that "the packet rate from this source is 50× higher than normal" and "the payload contains a known malicious byte sequence", which together contributed most to the model's anomaly score[33]. Research prototypes have applied methods like LIME and SHAP to IDS models on IoT networks, successfully retrieving explanations for black-box model results and thereby "increasing [the system's] interpretability" without significantly slowing it down[34]. These explanations are integrated into the SOC analyst's interface: when an alert is shown, the analyst sees a short explanation such as "Unusual DNS traffic volume and blacklisted domain detected". This was deliberately chosen to mimic the style of rule-based systems that analysts are familiar with – essentially providing a human-readable rationale even though the detection was done by a complex ML model. The impact has been significant on trust and efficiency: analysts report they are more likely to investigate an AI-generated alert if it comes with clear supporting evidence (as opposed to a cryptic anomaly score). It also speeds up the investigation – the analyst knows exactly what suspicious factors to look at (ports, IPs, frequencies) from the explanation, reducing the time to triage the alert.

In the realm of fraud detection (a cross-over of finance and security), XAI has similarly

[33] Diogo Gaspar, Paulo Silva, and Catarina Silva. "Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron". In: *IEEE Access* (2024).

[34] Omer Subasi et al. "A critical assessment of interpretable and explainable machine learning for intrusion detection". In: *arXiv preprint arXiv:2407.04009* (2024).

proven its worth. A notable case is Transmit Security, a company providing AI-driven fraud detection and identity threat prevention services. They integrated SHAP explanations into their detection engine to help clients understand each alert. According to Transmit Security, one challenge in fraud detection is that the systems often output a risk score or a yes/no decision with no context, making it hard for fraud investigators to act or trust the system [27]. By leveraging SHAP, their service can break down a fraud score into contributions of various features for each individual transaction or user session. For example, if a user's login is flagged as a suspected bot, the system might explain: "Login flagged due to: abnormal time of access (+0.3 risk), device not seen before (+0.4), failed security question (+0.2). Legitimate factors: password correct (-0.1)." The plus/minus values indicate how much each factor raised or lowered the risk. Such granular explanations help fraud teams in multiple ways: (1) Justification – they can confidently shut down a transaction and have detailed reasons on record, which is useful if a customer or auditor later questions it; (2) Model refinement – if certain features are consistently contributing to false positives, developers can tweak the model or data. Indeed, Transmit's engineers[27] use the collected SHAP outputs to identify when the model might be relying too heavily on a single factor and adjust the training accordingly. Importantly, the presence of explanations can aid in regulatory compliance for sectors like banking: to comply with anti-fraud regulations and privacy laws, companies must often document why an account was suspended or a transaction blocked, and XAI provides that documentation in a systematic way.

Cybersecurity XAI is also being applied in user behavior analytics and insider threat detection. These systems monitor patterns of user activity (logins, file access, etc.) to detect when a legitimate user's account might be compromised or a disgruntled insider might be abusing privileges. When an alert is raised, it is crucial to explain it to a security officer to decide on a response. For example, an insider threat model might raise an alert: "User X is 90% likely to be a security risk." An explanation engine can convert that into: "Unusual behavior detected: User X downloaded 5GB of data (10× typical) outside business hours and used an admin-only database command." Here we see a combination of statistical anomaly features (volume of data, time of action) and rule-based triggers (use of an admin command) that contributed to the model's prediction. Many deployed enterprise security tools now offer this kind of explanation report. It was integrated because early feedback from SOCs indicated that without at least a short description of why an alert was triggered, AI-based systems were often turned off or ignored. The explanation techniques used range from simple decision trees (approximating the behavior of a complex model) to more sophisticated natural language generators that translate model output into sentences. The integration challenges included making sure the explanations updated in real-time and did not overwhelm analysts with too much information. In practice, vendors found a balance by highlighting the top 3 factors in an alert. The result has been improved analyst confidence and faster incident response.For instance, Vodafone, in collaboration with Nokia, reported that after integrating machine learning and explainable AI into their anomaly detection system, the average validation time per network alert decreased significantly. Their new system automatically identifies abnormal network behavior across vast volumes of mobile traffic data. By combining real-time detection with interpretable explanations, analysts were able to quickly understand why an anomaly was flagged — such as unusual spikes in bandwidth usage or unexpected access patterns — and respond accordingly. Vodafone anticipates that up to 80% of anomalous mobile network issues and capacity demands can now be detected and handled automatically.[35]

A unique consideration in cybersecurity is that explainability may also have a downside: revealing how the detection works could potentially help adversaries circumvent it. This is a point of debate in the security community. Some argue that too much transparency (especially if explanations were exposed to end-users or attackers) might let hackers reverse-engineer the system. However, in the context of internal SOC tools, this is less of a concern since the explanations are for analysts' eyes only. In fact, some experts say that explainable AI can enhance security by allowing human experts to collaborate with AI and catch adversarial tricks. For example, if an attacker finds a way to fool the AI (say by shaping traffic patterns), a human analyst might notice something off in the explanations (like an irrelevant feature being the reason) and realize the model is being evaded. In one reported case, a security team noticed that their malware detection AI started giving explanations that didn't align with known malware indicators; this tipped them off that malware authors had adapted to exploit a blind spot, leading them to

[35] Vodafone Group PLC and Nokia. *Automating Anomaly Detection with Machine Learning in Telecom Networks*. https://www.acceldata.io/blog/automate-data-anomaly-detection-with-machine-learning-in-telecom-networks. 2023.

retrain the model. Such stories highlight a synergistic view: XAI can act as a debugging and monitoring tool, not just for developers but for end-users of security AI, ensuring the models continue to operate under the oversight of human logic and domain knowledge.

From a compliance and governance standpoint, industries like defense and critical infrastructure (which use AI for cybersecurity) have adopted XAI as part of meeting requirements for algorithmic accountability. Government standards (e.g. NIST's guidance on AI in security) emphasize the need for explanations in automated threat responses, to allow for audits and after-action reviews. If an AI automatically shuts down a network segment due to perceived intrusion, the operators need a log of why it took that action – akin to a flight recorder. XAI provides that log in an interpretable form: "Action taken because intrusion model output 0.95 (threshold 0.9) due to factors A, B, C." This level of detail is invaluable for post-mortem analysis and continuously improving both the AI and the incident response procedures.

In summary, XAI in cybersecurity is becoming a standard feature for any AI-driven solution. By employing techniques like feature attribution, rule extraction, and counterfactual analysis, these systems ensure that human analysts are kept in the loop and can understand the context of threats. The impact has been to greatly increase the trustworthiness of AI in a domain where trust is hard-won. Early anecdotal evidence and emerging studies suggest that security teams with explainable AI have higher intervention rates (meaning fewer missed real incidents) and lower false positive fatigue, as compared to those with "black-box" security AI. As cyber threats continue to evolve, the transparency provided by XAI may also become a strategic advantage – allowing defenders to adapt along with the AI. Nonetheless, ongoing research in adversarial machine learning reminds us that explanations must be used carefully. Just as in other domains, there is a balance to strike between openness and security. But the consensus so far is that the benefits of explainability – in terms of human-AI collaboration – outweigh the risks in the context of internal cybersecurity operations. By illuminating how AI systems make decisions about attacks and anomalies, XAI helps ensure that these decisions are sound, justifiable, and can be trusted under pressure, which is ultimately what matters most in cybersecurity scenarios.

Introducing B2B Matchmaking in Enterprise Contexts

Beyond the high-impact domains already discussed, explainability increasingly matters for inter-organizational collaboration. A particularly relevant setting is Business-to-Business (B2B) matchmaking, where platforms must surface and rank potential partners across large and heterogeneous candidate pools. Unlike consumer recommendations, B2B pairing is about bilateral compatibility—strategic, operational, and sometimes cultural—and decisions often leave an auditable trail. Some companies that operate in this context are: Grip operates an AI-driven event networking platform that produces matchmaking at very large scale, combining multiple machine-learning models with interaction data collected across events. In practice, it delivers ranked suggestions, facilitates meeting scheduling, and supports these flows through a widely used mobile application for conferences and trade shows.[36] While Brella provides intent-based matchmaking for events. Attendees define interests and goals; the system ingests these signals together with prior interactions to propose 1:1 meetings and populate meeting slots for in-person, hybrid, or virtual formats. In other words, networking is operationalized as a data-driven process that adapts to stated objectives and observed behavior.[37] On the other hand the Enterprise Europe Network (EEN) maintains a pan-European database of partnering opportunities where SMEs publish and search profiles for distribution, manufacturing, technology transfer, or R&D collaboration. Alongside the online listings, regional partners actively broker contacts and organize brokerage events to support cross-border matches.[38] Together, these cases illustrate current practice across contexts: private event-tech plat-

[36] Grip Events. *Create more valuable B2B events with smarter matchmaking.* Claims 70M+ yearly recommendations, 16 ML algorithms, and use of platform interactions; accessed 2025-09-18. 2025. URL: `https://www.grip.events/products/event-matchmaking`; Grip Events. *Grip - The AI-powered Event Platform Built for Business.* Features include AI matchmaking, meeting management, mobile app; accessed 2025-09-18. 2025. URL: `https://www.grip.events/`; Grip Events. *How to improve your event networking with AI matchmaking.* Explains app onboarding and data use to refine matches; accessed 2025-09-18. 2024. URL: `https://www.grip.events/news/how-to-improve-your-event-networking-with-ai-matchmaking`.

[37] Brella. *The most advanced event matchmaking software.* Product overview: AI-powered, intent-based matchmaking, 1:1 meeting proposals; accessed 2025-09-18. 2025. URL: `https://www.brella.io/event-matchmaking`; Brella Help Center. *Introduction — Brella Matchmaking.* How it works: attendees select interests/intents; AI recommendations; accessed 2025-09-18. 2025. URL: `https://help-organizers.brella.io/en/articles/177659-introduction`.

[38] Enterprise Europe Network. *About the Enterprise Europe Network.* States largest online database of business opportunities and partnering services; accessed 2025-09-18. 2025. URL: `https://een.ec.europa.eu/about-enterprise-europe-network`; Enterprise Europe Network. *Partnering opportunities.* Live partnering listings, searchable profiles; accessed 2025-09-18. 2025. URL: `https://een.ec.europa.eu/partnering-opportunities`.

forms prioritize volume, personalization, and meeting orchestration based on learning from profiles and interactions, whereas a public or semipublic infrastructure like EEN emphasizes discoverability, brokerage, and procedural clarity for international partnering.[39]

1.6.5 Conclusion

Across healthcare, finance, human resources, and cybersecurity, we find that Explainable AI has transitioned from a research concept to a practical necessity. In each sector, XAI techniques were tailored to the domain's needs: saliency maps and interpretable models in healthcare to align with clinical reasoning, SHAP values and reason codes in finance to satisfy regulatory scrutiny, counterfactual explanations in HR to provide actionable feedback and detect bias, and feature attributions in cybersecurity to assist analysts in understanding threats. These implementations show that context matters – the choice of XAI method was driven by what users (doctors, risk managers, HR officers, SOC analysts) would find intuitive and useful. Integration of XAI into existing systems often required innovation (as seen with GPU-accelerated SHAP for speed, or natural-language generation of HR feedback for clarity), but the payoff was a higher level of trust and user adoption of AI. We also observed common themes: XAI improved accountability (organizations can now explain and defend AI-driven decisions), aided in compliance with laws and ethical norms, and in some cases even enhanced model accuracy and fairness by exposing weaknesses (like spurious correlations or biases that were then fixed). Against this backdrop, B2B matchmaking emerges as a pertinent next case: contemporary platforms already operationalize large-scale, learning-based pairing in private event-tech settings (e.g., intent- and interaction-driven meeting suggestions) and, in parallel, maintain searchable partnering infrastructures in public or semi-public contexts with brokerage support and procedural clarity. These contrasting settings make explainability needs concrete—ranking drivers and proposal logic on one side; traceable selection rules and auditable criteria on the other—providing a practical testbed for assessing how explanations interact with privacy constraints and model performance. The next chapter therefore examines B2B matchmaking methodologies and actor types through this lens,

[39] Grip Events, *Create more valuable B2B events with smarter matchmaking*, see n. 36; Brella, see n. 37; Enterprise Europe Network, *About the Enterprise Europe Network*, see n. 38.

to specify where explanations are operationally required and how they can be integrated without undermining utility.

2

Methodologies for B2B Matchmaking in Digital Business Ecosystems

2.1 Introduction to B2B Matchmaking

2.1.1 Definition and Context

Business-to-Business (B2B) matchmaking represents a critical strategic process in modern digital economies, encompassing the systematic identification, evaluation, and connection of compatible business entities for mutually beneficial partnerships. Unlike traditional networking approaches that rely on fortuitous encounters or limited personal networks, contemporary B2B matchmaking employs structured methodologies to facilitate optimal business relationships across various dimensions including supply chain partnerships, distribution agreements, strategic alliances, and collaborative innovation ventures[1].

The emergence of Digital Business Ecosystems (DBEs) has fundamentally transformed the landscape of B2B interactions. As noted by[2], these ecosystems represent complex networks where organizations from diverse market segments converge on digital platforms to exchange value, share resources, and co-create innovations. Within these ecosystems, the challenge of identifying suitable partners has grown exponentially due to the sheer volume of potential connections, the diversity of business models, and the dynamic nature of market relationships.

The distinction between matchmaking and recommendation systems is particularly crucial in the B2B context. While recommendation systems suggest potential partners

[1] Mustapha Kamal Benramdane et al. "Supervised Machine Learning for Matchmaking in Digital Business Ecosystems and Platforms". In: *Information Systems Frontiers* 26.4 (2024), pp. 1331–1343. DOI: 10.1007/s10796-022-10357-3.

[2] Marco Iansiti and Roy Levien. *The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability.* Harvard Business School Press, 2004.

based on similarity metrics or historical patterns, matchmaking systems actively identify and evaluate compatibility across multiple dimensions, considering not only surface-level attributes but also strategic alignment, operational compatibility, and cultural fit. This approach is essential in B2B contexts where partnership decisions carry significant strategic implications and require substantial resource commitments [1].

2.1.2 Historical Evolution

The evolution of B2B matchmaking can be traced through three distinct phases, each reflecting broader technological and organizational paradigms. The pre-digital era (prior to1990s) was characterized by relationship-based networking, where business partnerships emerged primarily through personal connections, industry associations, and trade shows. Geographic proximity and sectoral boundaries strongly influenced partnership formation, with limited visibility beyond immediate business networks[3].

The digital transformation phase (1990s-2010s) witnessed the emergence of online B2B marketplaces and electronic data interchange systems. Platforms like Alibaba and ThomasNet democratized access to global supplier networks, while Customer Relationship Management (CRM) systems began capturing and structuring business relationship data. However, these systems primarily facilitated transactions rather than strategic partnerships, focusing on catalog-based matching rather than comprehensive compatibility assessment [3].

The current ecosystem era (2010s-present) represents a paradigm shift toward platform-based business models and ecosystem orchestration. Digital platforms now serve as intermediaries that not only connect businesses but also provide the infrastructure for value creation and exchange. The complexity of modern supply chains, the rise of Industry 4.0, and the increasing importance of data-driven decision-making have necessitated more sophisticated approaches to B2B matchmaking that can account for multifaceted compatibility criteria and dynamic market conditions [3].

[3] Mark de Reuver, Carsten Sørensen, and Rahul C. Basole. “The digital platform: a research agenda”. In: *Journal of Information Technology* 33 (2018), pp. 124–135.

2.1.3 Importance in Business Service Contexts

The strategic importance of effective B2B matchmaking in contemporary business environments cannot be overstated. For Small and Medium Enterprises (SMEs), which often lack the resources for extensive market research and partner discovery, structured matchmaking methodologies provide access to partnership opportunities that would otherwise remain invisible. Research by[4] identifies that the primary barriers to SME collaboration in digital environments include information asymmetry, trust deficits, and the complexity of evaluating partnership potential—all challenges that systematic matchmaking approaches can address.

In the context of business services, matchmaking plays a particularly crucial role in several dimensions. First, it enables service providers to identify clients whose needs align with their capabilities, reducing customer acquisition costs and improving service delivery outcomes. Second, it facilitates the formation of complementary service partnerships, where providers with different specializations can collaborate to deliver comprehensive solutions. Third, it supports the development of service ecosystems where multiple providers coordinate to create integrated value propositions[5].

The impact of effective matchmaking extends beyond individual transactions to influence broader market dynamics. By reducing search costs and information asymmetries, matchmaking systems increase market efficiency and enable more optimal resource allocation. They also facilitate innovation by connecting organizations with complementary capabilities, enabling knowledge transfer and collaborative development. Furthermore, in global markets, matchmaking systems help overcome geographic and cultural barriers, enabling cross-border partnerships that drive international trade and economic integration [5].

2.1.4 Overview of Methodological Approaches

The landscape of B2B matchmaking methodologies encompasses a spectrum from highly quantitative, algorithm-driven approaches to qualitative, judgment-based frameworks. While technological advances have enabled sophisticated computational matching

[4] Nikolai Kazantsev et al. “Investigating barriers to demand-driven SME collaboration in low-volume high-variability manufacturing”. In: *Supply Chain Management: An International Journal* 27.2 (2022), pp. 265–282.

[5] Iansiti and Levien, see n. 2, ch. "Integration, Innovation, and Adaptation".

using machine learning and artificial intelligence, the complexity of B2B relationships and the importance of tacit knowledge mean that qualitative methodologies remain essential, particularly in contexts where human judgment, strategic considerations, and cultural factors play decisive roles .

Quantitative approaches typically employ data-driven algorithms that analyze structured attributes such as company size, industry classification, financial metrics, and transaction histories. These methods excel at processing large volumes of potential matches and identifying patterns that might not be apparent to human observers. However, they often struggle to capture the nuanced factors that determine partnership success, such as organizational culture, leadership vision, and strategic intent .

Qualitative methodologies, by contrast, leverage human expertise, industry knowledge, and strategic reasoning to evaluate partnership potential. These approaches recognize that successful B2B relationships depend not only on objective compatibility metrics but also on subjective factors such as trust, communication styles, and shared values. They accommodate the ambiguity and complexity inherent in strategic decision-making, allowing for the integration of contextual knowledge and the consideration of factors that may not be easily quantifiable .

2.1.5 Scope and Objectives

The following analysis is dedicated to a representative selection of qualitative B2B matchmaking methodologies that have demonstrated effectiveness in applied contexts. The primary objective is to establish a framework that enables the selection and adaptation of the most suitable approach based on specific organizational needs and partnership goals.

The scope of this study encompasses methodologies applicable across various industries and partnership types, from supplier relationships to strategic alliances. While acknowledging the growing importance of data-driven techniques, the focus remains on methodologies where human judgment and strategic reasoning are paramount, recognizing these elements as irreplaceable in complex B2B decision-making. To facilitate a comparative assessment, each methodology is presented in terms of its theoretical underpinnings, implementation processes, and principal limitations, thereby highlighting its optimal application context.

2.1.6 Methodology 1: Weighted Scoring Matrix

Theoretical Principles

The Weighted Scoring Matrix represents one of the most fundamental yet powerful approaches to B2B matchmaking, rooted in multiattribute decision theory and compensatory decision-making models[6]. This methodology operates on the principle that partnership compatibility can be decomposed into discrete, measurable attributes, each contributing differentially to overall compatibility based on its relative importance. The theoretical foundation draws from utility theory, which posits that decisionmakers can assign numerical values to represent preferences and that overall utility can be calculated as a weighted sum of individual attribute utilities[7].

The compensatory nature of this approach allows strengths in certain areas to offset weaknesses in others, reflecting the reality that perfect matches rarely exist in B2B contexts. This characteristic distinguishes the Weighted Scoring Matrix from non-compensatory models that employ strict cutoff criteria. The methodology assumes that decision-makers can articulate their preferences explicitly and that these preferences remain relatively stable throughout the evaluation process [6].

From a cognitive perspective, the Weighted Scoring Matrix serves as a decision aid that structures complex evaluations into manageable components. By decomposing the matchmaking challenge into discrete criteria, it reduces cognitive load and enables systematic comparison across multiple potential partners. This structured approach helps overcome common decision-making biases such as the halo effect, where positive impressions in one area disproportionately influence overall assessment, and availability bias, where recent or memorable information receives disproportionate weight.

The mathematical foundation of the Weighted Scoring Matrix can be expressed as:

$$S = \sum_i (w_i \times r_i)$$

Where S represents the total score for a potential partner, wi represents the weight assigned to criterion i, and ri represents the rating or score for criterion i. This sim-

[6] Valerie Belton and Theodor J. Stewart. *Multiple Criteria Decision Analysis: An Integrated Approach.* Boston: Springer Science & Business Media, 2002.

[7] Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs.* Cambridge University Press, 1993.

ple yet effective formula enables transparent calculation and easy sensitivity analysis to understand how changes in weights or scores affect overall rankings.

Implementation Process

The implementation of a Weighted Scoring Matrix for B2B matchmaking follows a systematic process that begins with criteria identification and culminates in partner selection. The first phase involves comprehensive stakeholder consultation to identify relevant evaluation criteria. These criteria should reflect both strategic objectives and operational requirements, encompassing dimensions such as financial stability, technical capabilities, market position, cultural alignment, and innovation potential. The selection of criteria requires careful balance between comprehensiveness and practicality, as excessive criteria can complicate the evaluation process without meaningfully improving decision quality [6].

The second phase focuses on weight assignment, where the relative importance of each criterion is quantified. Various techniques can be employed for weight determination, including direct rating, point allocation, and pairwise comparison methods such as the Analytic Hierarchy Process (AHP)[8]. The choice of weighting method depends on the number of criteria, the availability of expert input, and the desired level of analytical rigor. It is crucial that weights reflect strategic priorities and that the weighting process involves key stakeholders to ensure buy-in and alignment.

The evaluation phase involves systematic assessment of potential partners against each criterion. This requires establishing clear scoring scales, typically ranging from 1-5 or 1-10, with explicit descriptors for each score level to ensure consistency across evaluators. Data collection may involve multiple sources including financial reports, capability assessments, reference checks, and direct interactions with potential partners. The scoring process should be documented thoroughly to enable review and revision if necessary.

The final phase involves score calculation and interpretation. Weighted scores are calculated by multiplying each criterion score by its weight and summing across all criteria. Although the numerical output produces a straightforward ranking of potential partners, the results should not be interpreted mechanically. Decision-makers need to examine how scores are distributed, how sensitive the ranking is to changes in weights, and whether

[8] Thomas L. Saaty. "Decision making with the analytic hierarchy process". In: *International Journal of Services Sciences* 1.1 (2008), pp. 83–98. DOI: 10.1504/IJSSCI.2008.017590.

there are qualitative aspects that the model does not capture. In this sense, the scores should serve as decision support tools rather than final, deterministic answers. [6].

Advantages and Limitations

The Weighted Scoring Matrix offers several compelling advantages in B2B matchmaking contexts. Its transparency and explicability make it particularly valuable in organizational settings where decisions must be justified to multiple stakeholders. The methodology provides a clear audit trail, documenting the rationale behind partnership decisions and enabling post-hoc analysis of decision quality. This transparency also facilitates organizational learning, as the criteria and weights can be refined based on partnership outcomes.

The flexibility of the approach represents another significant advantage. Organizations can customize criteria and weights to reflect their unique contexts and strategic priorities. The methodology can accommodate both quantitative metrics and qualitative assessments, enabling integration of hard data with expert judgment. Furthermore, the approach scales effectively from simple bilateral partnerships to complex multi-party alliances, with criteria adjusted to reflect the additional complexity.

The structured nature of the Weighted Scoring Matrix promotes consistency in evaluation across multiple potential partners and over time. By establishing explicit criteria and scoring guidelines, the methodology reduces the influence of individual biases and ensures that all candidates receive fair consideration. This consistency is particularly valuable in regulated industries or public sector contexts where procurement processes must demonstrate objectivity and fairness.

However, the methodology also presents notable limitations. The assumption of criteria independence often proves problematic in practice, as many partnership attributes are interrelated. For example, technical capabilities and innovation potential may be strongly correlated, leading to double-counting if both are included as separate criteria. The compensatory nature of the model may also be inappropriate in situations where certain criteria represent non-negotiable requirements [6].

The quality of outcomes depends heavily on the validity of weights and scores, both of which involve subjective judgment. Weight elicitation can be challenging, particularly when stakeholders have conflicting priorities or when preferences are context-dependent.

Similarly, scoring reliability may vary across evaluators, particularly for qualitative criteria where assessment standards may differ. These challenges necessitate careful process design and may require calibration exercises to ensure consistency [6].

Practical B2B Example

Consider a manufacturing company seeking to identify strategic suppliers for a new product line requiring specialized components. The company implements a Weighted Scoring Matrix with the following criteria and weights: Technical Capability (25%), Quality Standards (20%), Financial Stability (15%),Production Capacity (15%), Geographic Proximity (10%), Innovation Track Record (10%), Sustainability Practices (5%).

The evaluation team, comprising representatives from engineering, procurement, quality, and operations, develops detailed scoring rubrics for each criterion. Technical Capability, for instance, is assessed based on certification levels, equipment sophistication, and demonstrated expertise in similar components. Financial Stability combines credit ratings, revenue trends, and dependency ratios. Each potential supplier undergoes comprehensive evaluation, with scores assigned based on documentary evidence, site visits, and capability demonstrations.

Three suppliers emerge as top candidates with scores of 8.2, 7.9, and 7.6 respectively. However, sensitivity analysis reveals that small changes in the weight assigned to Geographic Proximity could alter the ranking, prompting deeper discussion about the true importance of supplier location. The company ultimately selects the second-ranked supplier after qualitative considerations reveal superior cultural alignment and collaboration potential not fully captured in the scoring model.

This example illustrates both the value and limitations of the Weighted Scoring Matrix. While the methodology provided structure and transparency to a complex decision, ultimate selection required integration of quantitative scores with qualitative judgment, demonstrating that the matrix serves as a decision aid rather than a decision replacement.

When to Use This Methodology

The Weighted Scoring Matrix is particularly appropriate when evaluation criteria can be clearly articulated and when stakeholders can reach consensus on relative importance. It works best in stable environments where criteria and weights remain valid throughout the evaluation period. Organizations with established procurement processes often find this methodology aligns well with existing procedures and governance requirements [6].

The approach is especially valuable when comparing a moderate number of potential partners (typically 5-20) across multiple dimensions. With fewer candidates, the overhead of developing the scoring framework may not be justified; with many more, the evaluation burden becomes excessive. The methodology also suits situations requiring decision documentation and justification, such as regulated industries or public sector procurement.

Organizations should consider alternative approaches when facing highly dynamic environments where partnership requirements evolve rapidly, when critical criteria are difficult to quantify or assess objectively, or when non-compensatory decision rules are more appropriate. The methodology may also prove inadequate when partnership success depends primarily on emergent properties of the relationship that cannot be predicted from individual partner attributes.

2.1.7 Methodology 2: Strategic Compatibility Framework

Theoretical Principles

The Strategic Compatibility Framework represents a holistic approach to B2B matchmaking that emphasizes alignment across multiple strategic dimensions rather than aggregation of individual attributes. Grounded in strategic management theory and the resource-based view of the firm[9], this methodology recognizes that successful partnerships emerge from complementary resources, aligned objectives, and compatible strategic trajectories. Unlike scoring-based approaches that treat compatibility as the sum of individual factors, the Strategic Compatibility Framework examines the systemic fit between organizations[10].

The theoretical foundation draws heavily from the concept of strategic fit, which suggests that partnership success depends on coherence between organizational strategies, structures, and cultures. This perspective acknowledges that organizations are complex

[9] Robert M. Grant. "Toward a Knowledge-Based Theory of the Firm". In: *Strategic Management Journal* 17.S2 (1996), pp. 109–122.

[10] Jeffrey H. Dyer and Harbir Singh. "The Relational View: Cooperative Strategy and Sources of Interorganizational Competitive Advantage". In: *Academy of Management Review* 23 (1998), pp. 660–679. URL: `https://api.semanticscholar.org/CorpusID:167965404`.

systems where individual elements interact in non-linear ways. Compatibility, therefore, cannot be reduced to a simple aggregation but must consider how different organizational characteristics combine to create synergies or conflicts.

The framework also incorporates insights from institutional theory, recognizing that organizations operate within broader institutional contexts that shape their behaviors and possibilities[11]. Partners operating under similar institutional pressures or within compatible institutional logics are more likely to develop successful relationships. This consideration extends beyond formal regulations to encompass industry norms, professional standards, and stakeholder expectations that influence organizational behavior.

Drawing from[12] value chain analysis, the framework examines how partners' value creation activities complement or conflict with each other. This analysis reveals opportunities for synergy where partners can leverage each other's strengths to create superior value propositions. It also identifies potential areas of conflict where overlapping capabilities or misaligned incentives might create tensions.

The dynamic capabilities perspective, introduced by[13], informs the framework's emphasis on strategic evolution and adaptation. Successful partnerships require not only current compatibility but also the ability to co-evolve as market conditions change. This dynamic view recognizes that strategic alignment is not a static condition but an ongoing process of mutual adjustment and learning.

Implementation Process

Implementation of the Strategic Compatibility Framework begins with comprehensive strategic profiling of both the organization and potential partners. This profiling extends beyond surface-level characteristics to examine deep structural elements including business models, value creation logic, competitive positioning, and growth trajectories. The profiling process typically employs structured interviews with senior leadership, analysis of strategic documents, and examination of historical strategic choices [10].

The strategic profiling phase utilizes multiple analytical tools to develop comprehen-

[11] Walter W. Powell. "Neither Market Nor Hierarchy: Network Forms of Organization". In: *Research in Organizational Behavior* 12 (1990), pp. 295–336.

[12] Michael E. Porter. *Competitive Advantage: Creating and Sustaining Superior Performance.* New York, NY: Free Press, 1985.

[13] David J. Teece, Gary Pisano, and Amy Shuen. "Dynamic capabilities and strategic management". In: *Strategic Management Journal* 18.7 (1997), pp. 509–533. DOI: 10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z.

sive organizational portraits. SWOT analysis identifies internal strengths and weaknesses alongside external opportunities and threats. Business model canvas mapping reveals how organizations create, deliver, and capture value. Core competency analysis identifies distinctive capabilities that provide competitive advantage. Resource mapping catalogs tangible and intangible assets that could be leveraged in partnerships [12].

The second phase involves mapping strategic dimensions and identifying areas of complementarity and conflict. Key dimensions typically include market positioning (e.g., cost leadership vs. differentiation), growth orientation (e.g., organic growth vs. acquisition), innovation approach (e.g., incremental vs. radical), and stakeholder priorities (e.g.,shareholder value vs. stakeholder balance). This mapping creates a multidimensional representation of strategic positioning that enables systematic comparison.

Strategic dimension mapping employs various visualization techniques to facilitate understanding and communication. Strategy maps illustrate cause-and-effect relationships between strategic objectives. Radar charts display organizational profiles across multiple dimensions simultaneously. Heat maps highlight areas of alignment and misalignment between potential partners. These visual tools support intuitive understanding of complex strategic relationships.

The compatibility assessment phase examines how organizational strategies interact across multiple scenarios and time horizons. This involves analyzing how partners' strategies might evolve and whether these evolutionary paths remain compatible. Scenario planning techniques help explore how compatibility might change under different environmental conditions. The assessment considers both current fit and dynamic alignment, recognizing that successful partnerships must accommodate strategic evolution.

The final phase involves synthesis and recommendation development. Unlike quantitative scoring methods, the Strategic Compatibility Framework produces qualitative assessments that describe the nature and implications of compatibility or incompatibility. Recommendations typically address not only partner selection but also partnership structure, governance mechanisms, and management approaches that can enhance compatibility or mitigate areas of misalignment [10].

Advantages and Limitations

The Strategic Compatibility Framework offers unique advantages in addressing the complexity of strategic partnerships. By examining systemic fit rather than isolated attributes, it captures interdependencies and emergent properties that simpler methodologies might miss. The framework's emphasis on strategic alignment helps identify partnerships with long-term potential rather than those offering only immediate tactical benefits. This strategic perspective is particularly valuable for partnerships intended to drive transformation or enter new markets [10].

The qualitative nature of the framework enables rich insights into partnership dynamics. Rather than reducing compatibility to a single number, the methodology provides nuanced understanding of where alignment exists, where tensions might arise, and how these might be managed. This depth of analysis supports more informed decision-making and better partnership design. The framework also accommodates the ambiguity and uncertainty inherent in strategic contexts, avoiding false precision that quantitative methods might suggest.

The framework's consideration of dynamic alignment represents a significant advantage in rapidly changing business environments. By examining not only current compatibility but also capacity for co-evolution, the methodology helps identify partnerships that can adapt and thrive amid uncertainty. This forward-looking perspective is particularly valuable for innovation partnerships or ventures into emerging markets where future conditions are difficult to predict [13].

However, the framework's complexity also presents challenges. The methodology requires significant expertise in strategic analysis and deep understanding of the industries and organizations involved. The qualitative nature of assessments can make comparison across multiple potential partners challenging and may introduce subjective bias. The demanding nature of strategic profiling often restricts the scope of partner evaluation, making it difficult to consider a broad pool of candidates, potentially causing organizations to overlook viable options [10].

The framework's emphasis on strategic alignment may underweight operational considerations that prove critical in partnership execution. Organizations with compelling strategic fit may still fail as partners due to incompatible operational processes, systems, or cultures. Additionally, the framework's focus on current and projected strategies may not adequately account for the adaptive capacity of organizations to adjust their strategies in response to partnership opportunities.

Practical B2B Example

A regional renewable energy company seeks partners to expand into international markets, requiring complementary capabilities in project development, financing, and local market access. The company employs the Strategic Compatibility Framework to evaluate three potential partners: a global infrastructure fund, an international engineering conglomerate, and a specialized renewable energy developer.

The strategic profiling reveals distinct strategic logics. The infrastructure fund pursues financial returns through portfolio diversification, with renewable energy representing one of multiple asset classes. The engineering conglomerate seeks to leverage renewable energy projects to showcase technical capabilities and secure long-term service contracts. The specialized developer focuses exclusively on renewable energy, pursuing market leadership through technical innovation and development expertise.

Compatibility assessment reveals that while the infrastructure fund offers substantial financial resources, its portfolio approach conflicts with the company's vision of deep market engagement and technical leadership. The engineering conglomerate presents operational synergies but potential conflicts arise from divergent views on project ownership and control. The specialized developer shows strong strategic alignment in vision and values but limited complementarity in capabilities, as both organizations possess similar strengths and weaknesses.

The analysis concludes that the engineering conglomerate offers the best strategic fit, provided partnership structures can be designed to accommodate different strategic priorities. The recommendation includes specific governance mechanisms to balance the conglomerate's service-oriented strategy with the company's development focus, such as separate joint ventures for development and operations phases. This example demonstrates how the Strategic Compatibility Framework provides nuanced insights that inform not just partner selection but partnership design.

When to Use This Methodology

The Strategic Compatibility Framework is most appropriate for high-stakes partnerships with significant strategic implications. These include joint ventures, strategic alliances, merger and acquisition targets, and long-term collaborative relationships that require deep integration. The methodology suits situations where partnership success depends on strategic alignment rather than tactical complementarity.

Organizations should employ this framework when evaluating a limited number of carefully pre-selected potential partners, as the resource intensity precludes broad screening applications. The methodology works best when sufficient information about potential partners' strategies is available or can be obtained through due diligence processes. It is particularly valuable in dynamic industries where strategic adaptation and co-evolution are critical for partnership success.

2.1.8 Methodology 3: Multi-Criteria Analysis (MCA)

Theoretical Principles

Multi-Criteria Analysis represents a sophisticated decision-making framework that explicitly addresses the multidimensional nature of B2B matchmaking while acknowledging the limitations of simple aggregation methods [6]. Unlike the Weighted Scoring Matrix, which assumes full compensability between criteria, MCA encompasses various techniques that can accommodate different types of criteria relationships, including non-compensatory rules, threshold effects, and criteria interactions. The theoretical foundation draws from decision science, operations research, and behavioral economics.

The methodology recognizes that B2B partnership decisions involve multiple, often conflicting objectives that cannot always be reduced to a single measure of utility. For instance, a potential partner might excel in technical capabilities but present financial risks, creating a decision dilemma that simple weighted averaging cannot adequately address. MCA provides frameworks for handling such trade-offs explicitly, enabling decision-makers to explore the implications of different priority structures and decision rules.

MCA approaches vary in their mathematical sophistication and underlying assumptions. Outranking methods like ELECTRE (Elimination and Choice Expressing Reality) build a preference relation by comparing alternatives pairwise, determining if one alternative "outranks" another based on the strength of the evidence, without needing to aggregate all criteria into a single score[14]. Similarly, PROMETHEE (Preference Ranking Organization Method for Enrichment Evaluation) also uses pairwise comparisons but calculates "preference flows" that quantify how much each alternative is preferred over all others, making the ranking explicit[15]. These methods recognize that not all criteria can be meaningfully combined and that certain performance differences may be too small to be significant.

Value-based methods like MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) take a different approach, using structured qualitative judgments from decision-makers (e.g., "the difference in attractiveness is 'strong'") and converting them into a numerical scale of scores, effectively building a quantitative model from qualitative input[16]. This approach bridges the gap between qualitative assessment and quantitative analysis, enabling decision-makers to express preferences in natural language while maintaining analytical rigor.

The theoretical elegance of MCA lies in its ability to handle different types of preference structures. Some criteria may follow linear preference functions where more is always better. Others may exhibit threshold effects where performance differences below a certain level are irrelevant. Still others may have ideal points where deviation in either direction reduces desirability. This flexibility enables more accurate representation of real decision preferences [6].

Implementation Process

The implementation of MCA for B2B matchmaking begins with problem structuring, where the decision context is carefully defined, stakeholders are identified, and the decision problem is framed. This phase is critical as it determines the scope of analysis and the types of criteria to be considered. Stakeholder analysis helps identify whose preferences should be incorporated and how different perspectives might be reconciled. The problem

[14] B. Roy. “The outranking approach and the foundations of ELECTRE methods”. In: *Theory and Decision* 31.1 (1991), pp. 49–73.

[15] J. P. Brans and P. Vincke. “A preference ranking organisation method: The PROMETHEE method for multiple criteria decision-making”. In: *Management Science* 31.6 (1985), pp. 647–656. DOI: 10.1287/mnsc.31.6.647.

[16] Carlos A. Bana e Costa and Jean-Claude Vansnick. “MACBETH: An interactive path towards the construction of cardinal value functions”. In: *International Transactions in Operational Research* 1.4 (1994), pp. 489–500.

structuring phase often reveals that what initially appears as a single decision actually involves multiple interconnected choices [6].

Problem structuring employs various techniques to ensure comprehensive problem definition. Cognitive mapping helps stakeholders articulate their understanding of the decision situation and identify relevant factors. Value-focused thinking shifts attention from alternatives to objectives, ensuring that fundamental goals drive the analysis rather than available options. Soft systems methodology addresses the social and political dimensions of the decision context.

Criteria development in MCA goes beyond simple identification to include careful specification of measurement scales, preference directions, and criteria types. Criteria may be quantitative or qualitative, and scales may be cardinal, ordinal, or nominal. The methodology distinguishes between different criteria roles: some serve as objectives to be optimized, others as constraints that must be satisfied, and still others as goals with specific target levels. This nuanced treatment of criteria enables more accurate representation of decision requirements.

The preference elicitation (i.e., the process of systematically collecting information about decision-makers' priorities) phase varies depending on the MCA technique used. It may involve directly rating the importance of criteria, comparing alternatives in pairs, or defining thresholds for indifference and preference. Modern MCA software often supports this process with interactive tools that guide decision-makers to express their preferences more consistently.

The analysis phase applies the selected MCA technique to generate insights about partner compatibility. This typically produces not a single "optimal" solution but rather a set of non-dominated alternatives and sensitivity analyses showing how recommendations change under different preference structures. The robustness of recommendations is assessed by examining stability across reasonable variations in parameters. The output includes not just rankings but also detailed performance profiles that show each alternative's strengths and weaknesses [6].

Advantages and Limitations

MCA offers several significant advantages for B2B matchmaking. The methodology's flexibility in handling different types of criteria and preference structures makes it applicable to a wide range of partnership contexts. The explicit treatment of trade-offs and the ability to incorporate non-compensatory decision rules provide more realistic modeling of actual decision processes. The methodology can handle incomplete information and uncertainty through techniques like interval judgments and fuzzy set theory [6].

The structured approach of MCA helps surface and resolve conflicts between different stakeholder perspectives. By making preferences and trade-offs explicit, the methodology facilitates negotiation and consensus building. The detailed performance profiles generated by MCA provide rich information for decision-making, going beyond simple rankings to show why certain alternatives perform well or poorly. This transparency supports both decision justification and organizational learning.

MCA's ability to handle criteria interactions represents a significant advantage over simpler methods. In B2B contexts, criteria often exhibit synergistic or antagonistic relationships. For example, technical innovation and cost-effectiveness may be negatively correlated, while quality standards and regulatory compliance may reinforce each other. MCA techniques can explicitly model these interactions, leading to more accurate assessment of partnership potential.

However, MCA also presents significant challenges. The complexity of some MCA techniques can be intimidating for practitioners without specialized training. The cognitive burden of providing consistent preference information across multiple criteria and alternatives can be substantial. The proliferation of MCA methods, each with different assumptions and requirements, can make method selection challenging. There is often a trade-off between methodological sophistication and practical implementability [6].

The quality of MCA results depends critically on the quality of input data and preference information. Elicitation biases, inconsistent judgments, and incomplete information can all compromise results. While MCA methods include consistency checking mechanisms, achieving fully consistent preference sets from multiple stakeholders remains challenging. The seeming objectivity of mathematical models can hide the fact that subjective choices—such as which criteria to include, how to measure them, and how to set preferences—are always part of the process.

Practical B2B Example

A pharmaceutical company seeks to identify contract research organizations (CROs) for clinical trial management across multiple therapeutic areas. The decision involves complex trade-offs between cost, quality, speed, geographic coverage, therapeutic expertise, and regulatory compliance. The company implements MCA using the PROMETHEE method, which allows for different preference functions for different criteria.

The criteria structure includes six main dimensions with multiple sub-criteria: Scientific Capabilities (therapeutic area expertise, protocol design experience, biostatistics capabilities), Operational Excellence (site activation speed, patient recruitment rates, data quality metrics), Geographic Reach (coverage of target markets, local regulatory knowledge), Financial Considerations (cost per patient, payment terms, financial stability), Quality Systems (regulatory inspection history, quality certifications, corrective action responsiveness), and Innovation Potential (technology platforms, adaptive trial design experience, real-world evidence capabilities).

For each criterion, the team specifies preference functions that reflect the nature of performance differences. For cost, a linear preference function indicates that any cost reduction is valuable. For regulatory compliance, a threshold function indicates that differences below a certain level are negligible, but larger differences matter significantly. For therapeutic expertise, a categorical preference function distinguishes between full expertise, partial expertise, and no expertise.

The analysis evaluates seven pre-qualified CROs, generating preference flows that indicate each CRO's strengths and weaknesses. The results reveal that no single CRO dominates across all criteria, but three emerge as part of the efficient frontier. Sensitivity analysis shows that the top recommendation remains stable across reasonable weight variations, but the ranking of second-tier alternatives depends significantly on the relative importance assigned to cost versus quality.

The company ultimately selects a portfolio approach, partnering with two CROs that show complementary strengths: one excelling in operational efficiency and cost-effectiveness for large, straightforward trials, another specializing in complex protocols requiring deep scientific expertise. This example illustrates how MCA supports sophisticated decision-making that goes beyond simple ranking to inform strategic partnership portfolio design.

When to Use This Methodology

MCA is particularly appropriate when partnership decisions involve multiple, potentially conflicting criteria that cannot be easily monetized or reduced to a common scale. The methodology suits complex decisions where different stakeholders hold different priorities and where these differences need to be explicitly addressed. MCA works well when decision-makers can invest time in careful problem structuring and preference articulation.

The methodology is valuable when the decision context requires detailed documentation of the rationale behind partnership choices, such as in regulated industries or public procurement. MCA's ability to handle different types of criteria makes it suitable for partnerships involving both quantitative performance metrics and qualitative strategic considerations. This approach is especially useful in situations where organizations must evaluate several potential partnerships at once, as part of a broader portfolio of decisions.

2.1.9 Methodology 4: Profiling and Gap Analysis

Theoretical Principles

Profiling and Gap Analysis represents a diagnostic approach to B2B matchmaking that focuses on identifying and evaluating differences between current and desired states. The methodology draws from competency modeling, organizational assessment, and strategic gap analysis traditions [10]. Rather than scoring or ranking potential partners, this approach creates detailed profiles that illuminate areas of alignment and misalignment, enabling informed decisions about partnership viability and requirements for success.

The theoretical foundation rests on the premise that successful partnerships require not perfect matches but rather complementary profiles where partners' strengths address each other's gaps. This perspective shifts the focus from finding similar organizations to identifying synergistic combinations. The methodology recognizes that gaps can represent both opportunities for value creation through complementarity and risks that require mitigation strategies.

The approach incorporates concepts from organizational capability theory, which distinguishes between different types of organizational resources and capabilities. Some gaps may involve easily transferable resources that partners can share readily, while others may involve deeply embedded capabilities that resist transfer. Understanding the nature of gaps helps predict whether partnerships can successfully bridge them and what mechanisms might be required.

Drawing from the knowledge-based view of the firm [9], the methodology recognizes that organizational knowledge exists at multiple levels - individual, group, and organizational - and in multiple forms - explicit and tacit. Gap analysis must therefore consider not only what knowledge or capabilities are missing but also how readily they can be acquired or accessed through partnership. This detailed understanding informs decisions about partnership structure and knowledge transfer mechanisms.

The dynamic capabilities framework [13] informs the methodology's treatment of capability gaps. Some gaps reflect missing operational capabilities that can be readily acquired or developed. Others reflect missing dynamic capabilities - the ability to sense opportunities, seize them, and reconfigure resources accordingly. Partnerships aimed at addressing dynamic capability gaps require different structures and management approaches than those addressing operational gaps.

Resource dependency theory[17] provides insights into how organizations can use partnerships to address critical resource gaps while managing the dependencies such partnerships create. The methodology examines not only what gaps exist but also the strategic importance of those gaps and the risks associated with different approaches to addressing them. This analysis helps organizations balance the benefits of accessing complementary resources against the costs of increased dependence.

Implementation Process

Implementation begins with comprehensive profiling of the focal organization to establish a baseline understanding of current capabilities, resources, and requirements. This self-assessment employs multiple data collection methods including capability audits, process mapping, resource inventories, and performance benchmarking. The profiling extends beyond current state to include strategic aspirations and identified improvement areas. Creating an honest, detailed self-profile requires overcoming organizational tendencies toward overestimation of capabilities and underestimation of weaknesses [10].

[17] Jeffrey Pfeffer and Gerald R. Salancik. *The external control of organizations: A resource dependence perspective.* Harper & Row, 1978.

The self-profiling phase utilizes various analytical tools to develop a comprehensive organizational portrait. Capability maturity models assess the sophistication of organizational processes across different domains. Value stream mapping identifies where value is created and where inefficiencies exist. Skills inventories catalog individual and collective competencies. Technology audits assess the sophistication and integration of technical systems. Cultural assessments examine organizational values, norms, and practices that influence partnership success.

The second phase involves defining ideal partner profiles based on strategic objectives and identified gaps. This is not a single profile but rather a set of profiles reflecting different partnership strategies. For instance, one profile might describe an ideal capability-complementing partner that fills specific technical gaps, while another might describe a market-access partner that provides geographic or segment reach. These ideal profiles serve as templates against which actual potential partners are assessed.

Ideal profile development requires careful consideration of partnership objectives and constraints. Strategic objectives determine what capabilities or resources partners should provide. Operational requirements specify how partners should be able to work together. Cultural preferences identify desired organizational characteristics that facilitate collaboration. Risk tolerance influences the acceptable level of partner dependency and the required partner stability. These considerations combine to create multidimensional ideal partner profiles.

The assessment phase involves detailed profiling of potential partners using available information sources. This includes public information, direct inquiries, reference checks, and, where possible, site visits or capability demonstrations. The challenge lies in obtaining accurate, comprehensive information about potential partners who may be reluctant to share detailed capability information before partnership agreements are in place. The assessment must distinguish between claimed and demonstrated capabilities.

Gap analysis compares actual partner profiles against ideal profiles and against the focal organization's profile to identify specific areas of alignment and misalignment. This analysis goes beyond simple gap identification to examine gap characteristics: Are gaps complementary or duplicative? Are they bridgeable through collaboration or require fundamental changes? What are the implications of leaving certain gaps unaddressed? The analysis produces detailed gap maps that visualize the partnership landscape.

The synthesis phase integrates insights from gap analysis to develop partnership recommendations. This includes not only partner selection but also partnership design recommendations that address identified gaps. For complementary gaps, the analysis specifies how partners can leverage each other's strengths. For problematic gaps, it identifies mitigation strategies such as capability development, third-party involvement, or alternative partnership structures [10].

Advantages and Limitations

Profiling and Gap Analysis offers unique advantages in providing detailed, actionable insights about partnership potential. The methodology's diagnostic nature helps organizations understand not just whether to partner but how to structure partnerships for success. By identifying specific gaps and complementarities, the approach informs decisions about partnership scope, resource allocation, and capability development priorities. The visual nature of profile comparisons and gap maps facilitates communication with stakeholders who may struggle with more abstract analyses.

The methodology's focus on complementarity rather than similarity helps identify subtle partnership opportunities. Organizations that appear quite different on surface dimensions may reveal deep complementarities when profiled systematically. This can lead to innovative partnerships that create significant value through synergy. The approach also helps organizations develop realistic expectations about what partnerships can and cannot achieve [10].

The detailed nature of gap analysis supports sophisticated partnership design. By understanding precisely what gaps exist and their characteristics, organizations can develop targeted strategies for addressing them. This might include specific capability transfer mechanisms, joint development programs, or complementary investment strategies. The methodology thus bridges the gap between partner selection and partnership implementation.

The methodology's emphasis on self-awareness represents both a strength and a requirement. Organizations that develop accurate self-profiles gain valuable insights that extend beyond partnership decisions to inform broader strategic planning. The process of self-profiling often reveals previously unrecognized strengths and weaknesses, enabling more informed strategic choices.

However, the methodology also faces significant limitations. The quality of analysis depends heavily on the accuracy and completeness of profiles, which can be challenging to develop, particularly for potential partners. Self-assessment biases can lead to unrealistic profiles of the focal organization, while information asymmetries limit understanding of potential partners. The static nature of profiles may not capture organizational dynamism and adaptive capacity.

The methodology's diagnostic focus may not provide clear decision rules for partner selection. While gap analysis reveals the landscape of possibilities, it does not necessarily indicate which gaps are most critical or which complementarities most valuable. Decision-makers must still exercise judgment in interpreting gap analysis results. The approach may also underemphasize relationship factors like trust, communication, and cultural fit that prove critical for partnership success but are difficult to profile objectively [10].

Practical B2B Example

A traditional automotive manufacturer seeks partners to develop electric vehicle (EV) capabilities while maintaining its core internal combustion engine business. The company conducts comprehensive self-profiling that reveals strong capabilities in manufacturing, supply chain management, dealer networks, and brand equity, but significant gaps in battery technology, power electronics, charging infrastructure knowledge, and software development for vehicle control systems.

The company develops three ideal partner profiles: a "Technology Partner" providing battery and power electronics expertise, a "Software Partner" delivering connected vehicle platforms and autonomous driving capabilities, and an "Infrastructure Partner" offering charging network access and energy management solutions. Each profile specifies required capabilities, resource commitments, and collaboration models.

Assessment of potential partners reveals interesting patterns. Traditional automotive suppliers offer incremental improvements but lack transformative EV technologies. Pure-play EV technology companies possess cutting-edge capabilities but lack automotive industry experience and scale manufacturing knowledge. Technology giants bring software expertise and capital but may seek control levels incompatible with the manufacturer's strategic autonomy requirements.

Gap analysis identifies a battery technology startup with breakthrough solid-state battery technology but lacking manufacturing capabilities and automotive application experience. The complementarity is nearly perfect: the manufacturer's gaps align precisely with the startup's strengths and vice versa. However, the analysis also reveals critical secondary gaps in areas like quality systems and supply chain integration that must be addressed.

The company structures a partnership that leverages the complementarity while addressing secondary gaps through specific initiatives. The partnership includes technology licensing, joint development programs, and capability transfer mechanisms. The manufacturer provides manufacturing expertise and industry knowledge while gaining access to breakthrough battery technology. Governance structures ensure both parties benefit from the partnership while maintaining their strategic independence.

This example shows how Profiling and Gap Analysis enables sophisticated partnership design based on detailed understanding of complementarities and gaps. The methodology's diagnostic insights inform not just partner selection but also partnership structure, governance, and implementation strategies.

When to Use This Methodology

Profiling and Gap Analysis is most appropriate when organizations seek partners to address specific capability gaps or when partnership strategy involves building complementary resource combinations. The methodology works well for technology partnerships, market entry alliances, and capability development initiatives where clear gaps exist between current and desired states [10].

The effectiveness of this methodology is predicated on several conditions. Firstly, it requires a candid organizational self-assessment, as unrealistic profiles can invalidate the analysis. Secondly, its success depends on access to sufficient information regarding potential partners, typically obtained through due diligence, industry knowledge, or preliminary collaboration. The framework proves most valuable when applied to partnerships aimed at capability transfer or joint development, as the gap analysis can precisely define the necessary transfer requirements and mechanisms.

2.1.10 Methodology 5: KPI-based Decision Matrix

Theoretical Principles

The KPI-based Decision Matrix methodology grounds B2B matchmaking in performance measurement theory and evidence-based management principles. Rather than relying on subjective assessments or strategic projections, this approach evaluates potential partners based on demonstrated performance across key performance indicators (KPIs) relevant to partnership objectives. The theoretical foundation combines performance management frameworks with decision theory to create a systematic, metrics-driven approach to partner selection[18].

The methodology rests on the premise that past performance, properly measured and contextualized, provides the best prediction of future partnership success. This empirical orientation distinguishes the KPI-based approach from methods that rely heavily on qualitative judgments or future projections. By anchoring evaluation in measurable outcomes, the methodology aims to reduce selection bias and increase decision objectivity.

This approach builds on the balanced scorecard framework [18], which emphasizes that performance is multidimensional and that stakeholders may prioritize different aspects of it. KPIs may span financial, operational, innovation, and sustainability dimensions, reflecting the complex value creation logic of modern B2B partnerships. The methodology acknowledges that KPI selection and interpretation require careful consideration of context, as the same metric may have different implications in different industries or strategic contexts.

Statistical learning theory informs the methodology's approach to predicting future performance from historical data. The relationship between past and future performance is modeled probabilistically, recognizing that while past performance provides valuable signals, it does not guarantee future outcomes. The methodology therefore incorporates uncertainty analysis and considers performance trends rather than point estimates.

The signaling theory perspective[19] suggests that observable performance metrics serve as signals of underlying organizational capabilities and characteristics. High performance

[18] Robert S. Kaplan and David P. Norton. "The balanced scorecard: Measures that drive performance". In: *Harvard Business Review* 70.1 (1992), pp. 71–79.

[19] Michael Spence. "Job Market Signaling". In: *The Quarterly Journal of Economics* 87.3 (1973), pp. 355–374.

on relevant KPIs signals not only specific achievements but also broader organizational competencies such as operational excellence, innovation capability, or market responsiveness. The methodology leverages these signals while recognizing their limitations and potential for manipulation.

Implementation Process

Implementation begins with KPI framework development, where relevant performance indicators are identified based on partnership objectives and critical success factors. This involves translating strategic partnership goals into measurable indicators that can be tracked and verified. The KPI selection process must balance comprehensiveness with practicality, focusing on indicators that truly differentiate partner performance and predict partnership success [18].

The KPI selection phase requires careful consideration of measurement validity and reliability. Validity ensures that KPIs actually measure what they purport to measure and relate meaningfully to partnership success. Reliability ensures that measurements are consistent and reproducible. The selection process also considers data availability, as the most theoretically relevant KPIs may not be practically measurable.

KPI categories typically include: • Financial Performance: Revenue growth, profitability margins, cash flow stability, return on assets, debt-to-equity ratios • Operational Excellence: On-time delivery rates, quality metrics, capacity utilization, cycle times, productivity measures • Innovation Capability: R&D investment intensity, patent portfolio, new product introduction rates, innovation pipeline value • Market Position: Market share, customer satisfaction scores, brand value, customer retention rates, market growth • Sustainability and Governance: ESG ratings, compliance records, employee satisfaction, safety performance, environmental impact metrics

The framework development includes establishing KPI definitions, measurement methods, and data sources. Standardization is critical to ensure comparability across potential partners. This may require developing common metrics that accommodate industry differences while maintaining meaningful comparability. The framework must also specify timeframes for performance measurement, as some KPIs may show high volatility while others change slowly.

Data collection represents a critical and often challenging phase. KPI data may come from various sources including financial reports, industry databases, regulatory filings, customer reviews, and direct partner disclosures. Data quality varies significantly across sources and partners, requiring careful validation and potentially adjustment for comparability. Missing data is common and must be handled systematically, whether through estimation, exclusion, or alternative metrics.

The decision matrix construction involves organizing KPI data into a structured format that enables systematic comparison. This includes normalization to account for scale differences, potentially weighting to reflect relative importance, and aggregation to produce summary assessments. Unlike simple scoring matrices, the KPI-based approach maintains traceability to underlying performance data, enabling drill-down analysis to understand score drivers.

Performance analysis goes beyond simple ranking to examine performance patterns and trends. Time series analysis reveals whether performance is improving or declining. Benchmarking against industry standards provides context for absolute performance levels. Correlation analysis examines relationships between different KPIs to identify underlying performance drivers. This comprehensive analysis provides rich insights into partner capabilities and potential [18].

Advantages and Limitations

The KPI-based Decision Matrix offers compelling advantages through its empirical grounding and objectivity. By basing evaluation on measurable performance metrics, the methodology reduces the influence of subjective biases and political considerations that can affect partnership decisions. The evidence-based approach provides strong justification for partner selection decisions, particularly important in organizations with formal governance requirements or stakeholder accountability [18].

The methodology's transparency enables clear communication about selection criteria and decision rationale. Stakeholders can understand exactly why certain partners were selected by examining their performance on specific KPIs. This transparency also facilitates post-decision learning, as organizations can track whether partners selected based on certain KPIs indeed deliver expected partnership value.

The quantitative nature of the approach enables sophisticated analyses including statistical modeling of KPI-outcome relationships, benchmarking against industry standards, and trend analysis to identify improving or declining performers. The methodology can be automated partially, enabling efficient screening of large numbers of potential partners and continuous monitoring of partner performance.

The scalability of the KPI-based approach represents a significant advantage. Once the KPI framework is established, evaluating additional potential partners requires relatively little incremental effort. This makes the methodology particularly suitable for situations involving many potential partners or ongoing partner selection needs. The approach also supports portfolio analysis, enabling organizations to optimize partnership portfolios based on complementary performance profiles.

However, the methodology faces significant limitations. The focus on historical performance may miss organizations with transformation potential or those whose past performance reflects different strategic contexts. KPIs capture what is measurable, not necessarily what is important, potentially overlooking critical soft factors like cultural fit, innovation potential, or collaborative capability. The availability and quality of KPI data varies significantly, potentially biasing selection toward larger, more transparent organizations [18].

The methodology risks promoting KPI manipulation, where potential partners focus on optimizing reported indicators rather than strengthening their underlying capabilities.The lag between performance and measurement means KPIs may not reflect current realities, particularly in dynamic industries. Creating meaningful, comparable KPIs across diverse organizations and industries remains challenging, often requiring compromises that reduce discriminatory power.

Practical B2B Example

A global logistics company seeks regional distribution partners to support e-commerce expansion in emerging markets. The company develops a KPI framework focusing on operational excellence, financial health, market presence, and service quality. Operational KPIs include on-time delivery rate (target: >95

Data collection reveals significant variations in data availability across regions and potential partners. Established logistics companies provide comprehensive performance data through annual reports and industry databases. Smaller regional specialists offer limited metrics, requiring alternative data sources such as customer references and operational audits. The company develops a tiered approach, with minimum data requirements for initial screening and progressively detailed requirements for shortlisted candidates.

The decision matrix analysis reveals interesting patterns. Some partners showing strong financial KPIs demonstrate weak operational performance, suggesting profitability through service compromises. Others excel operationally but show concerning financial trends, raising sustainability questions. A cluster of mid-sized regional specialists shows balanced performance across dimensions, though at lower absolute levels than larger competitors.

Deeper analysis reveals that KPI performance correlates strongly with local market characteristics. Partners in mature markets show higher efficiency metrics but lower growth rates, while those in emerging markets demonstrate opposite patterns. This insight leads to portfolio partnership strategy, selecting partners whose KPI profiles align with specific market requirements rather than seeking uniformly high performers.

The company ultimately selects three regional partners based on differentiated KPI profiles. For mature European markets, they choose a partner with exceptional operational efficiency (98.5

This example illustrates how KPI-based Decision Matrix enables nuanced, context-sensitive partner selection based on empirical performance evidence. The methodology's data-driven insights inform not just partner selection but also performance expectations and partnership management strategies.

When to Use This Methodology

The KPI-based Decision Matrix is most appropriate when reliable performance data is available for potential partners and when past performance meaningfully predicts future partnership success. The methodology suits industries with standardized performance metrics and reporting requirements, such as logistics, manufacturing, and financial services. It works well for operational partnerships where execution excellence is paramount [18].

The approach is valuable when organizations need to justify partner selection decisions with objective evidence, such as in regulated industries or public procurement contexts. It suits situations where large numbers of potential partners must be screened efficiently, as KPI-based filtering can be partially automated. The methodology also works well for ongoing partner performance management, as the same KPIs used for selection can monitor partnership execution.

Organizations should consider alternative approaches when entering new markets or industries where historical performance may not predict future success, when innovation or transformation is the primary partnership objective, or when critical success factors are difficult to measure quantitatively. The methodology may be less suitable for strategic partnerships where soft factors like vision alignment and cultural fit dominate. In industries undergoing disruption, historical KPIs may mislead rather than inform.

Table 2.1: B2B Matchmaking – Methodology Summary

Methodology	**Brief Description**	**Core Theoretical Principles**	**Key Advantages**	**Key Limitations**
Weighted Scoring Matrix	Evaluation via weighted attributes with trade-offs.	MAUT; compensatory models; utility theory.	Transparent; explainable; flexible; consistent.	Criteria interdependence; compensatory nature not always suitable; subjective weights.
Strategic Compatibility Framework	Assesses strategic alignment and systemic fit for long-term partnerships.	RBV; strategic fit; dynamic capabilities.	Handles complexity; captures interdependencies; alignment over time.	Complex and resource-intensive; hard to compare; may underweight operations.
Multi-Criteria Analysis (MCA)	Decision framework for multiple, conflicting objectives; can be non-compensatory.	Outranking (ELECTRE, PROMETHEE); value-based methods (e.g., MACBETH); decision science.	Flexible criteria/preferences; explicit trade-offs; detailed performance profiles; handles interactions.	Methodologically complex; heavy preference elicitation; results depend on inputs/bias in elicitation.
Profiling and Gap Analysis	Diagnoses differences between current and target states to define requirements.	Gap analysis; organizational capabilities; knowledge-based view; resource dependence.	Actionable insights; identifies gaps/complementarities; guides partnership design.	Depends on profile accuracy; self-assessment bias; no explicit decision rule.
KPI-based Decision Matrix	Partner evaluation based on KPIs aligned with objectives.	Performance measurement; evidence-based management; statistical learning; signaling theory.	Empirical and objective; reduces subjective bias; transparent; scalable; supports portfolio analysis.	Backward-looking; misses soft factors; KPI gaming risk; data quality variability.

2.2 Comparative Analysis

2.2.1 Comparative Framework

The five qualitative methodologies presented in this chapter each offer distinct approaches to the B2B matchmaking challenge, with different theoretical foundations, implementation requirements, and optimal application contexts. To facilitate informed methodology selection, this comparative analysis examines the methodologies across multiple dimensions critical to practical implementation and effectiveness.

The comparison employs evaluation criteria relevant to organizational decision-makers: implementation complexity, resource requirements, scalability potential, transparency of process, and adaptability to different contexts. Implementation complexity encompasses both technical sophistication and cognitive demands placed on users. Resource requirements include time, expertise, and data needs. Scalability addresses the methodology's ability to handle varying numbers of potential partners. Process transparency relates to the explicability of decisions to stakeholders. Adaptability concerns the methodology's flexibility across different industries and partnership types.

Additionally, the analysis considers the robustness of the methodology, stakeholder acceptance, and the potential for integration with existing organizational processes.These factors significantly influence the practical viability of methodology adoption and long-term sustainability within organizational contexts.

2.2.2 Detailed Comparison

Implementation Complexity and Resource Requirements

The methodologies exhibit a wide spectrum of complexity and resource requirements, ranging from conceptually intuitive frameworks to those demanding specialized expertise. At the lower end of this spectrum, the Weighted Scoring Matrix and the KPI-based Decision Matrix present a low-to-medium complexity. Their conceptual foundations are straightforward and easily understood by business practitioners. However, this apparent simplicity can be deceptive, as their effectiveness hinges on careful design choices in criteria selection, determination of weights, and data validation, which require moderate resource investments, primarily in terms of stakeholder time and data collection [6].

In contrast, the Strategic Compatibility Framework represents the highest level of complexity and resource intensity. It demands sophisticated strategic analysis capabilities, deep industry knowledge, and significant involvement from senior management. The qualitative and multifaceted nature of this framework often requires skilled facilitators to synthesize complex information, making it inherently resource-intensive in both time and expertise.Occupying a variable middle ground are Multi-Criteria Analysis (MCA) and Profiling and Gap Analysis. The complexity and resource needs of MCA depend heavily on the specific technique employed; simple weighted-sum models are comparable to scoring matrices, while sophisticated outranking methods (e.g., ELECTRE, PROMETHEE) require specialized knowledge and software. Similarly, Profiling and Gap Analysis, while conceptually straightforward, demands significant resources for the comprehensive data gathering required to build accurate organizational profiles .

Scalability

Scalability, defined as the ability to handle a growing number of potential partners, varies significantly across the methodologies. The KPI-based Decision Matrix and the Weighted Scoring Matrix demonstrate the highest scalability. Once their respective frameworks are established, additional partners can be evaluated with consistent and relatively low incremental effort, making them suitable for screening large pools of candidates (e.g., 20-50+).

Conversely, the Strategic Compatibility Framework exhibits low scalability due to the resource-intensive nature of in-depth strategic profiling, which typically limits its application to a small number of carefully pre-selected partners (e.g., 3-5). Attempts to scale this approach beyond such a number would likely compromise the depth and quality of the analysis. Profiling and Gap Analysis and Multi-Criteria Analysis offer medium scalability. The former is constrained by the effort of profiling and is typically suitable for 5-10 partners, while the latter can effectively handle 10-15 partners with the aid of computational tools, though its limiting factor often becomes the cognitive burden of preference elicitation rather than computation itself .

Process Transparency

Transparency, or the explicability of the decision-making process, is a critical factor for stakeholder buy-in. Methodologies grounded in explicit calculation, such as the Weighted Scoring Matrix and the KPI-based Decision Matrix, offer the highest degree of transparency. Their use of clear criteria, weights, and scores creates a discernible audit trail and allows stakeholders to readily understand the decision rationale [6].

The other methodologies offer more differentiated forms of transparency. Multi-Criteria Analysis is transparent in its methodological procedures, but the sophistication of certain techniques can obscure the underlying analysis for non-technical stakeholders. Profiling and Gap Analysis achieves transparency through intuitive visual tools like gap maps and profile comparisons. Finally, the Strategic Compatibility Framework provides medium transparency; while its analytical logic is clear, its conclusions are qualitative and subjective, meaning their clarity depends heavily on the communication skills of the analysts.

Adaptability

The adaptability of a methodology to different contexts, industries, and partnership types is crucial for its practical utility. The Weighted Scoring Matrix and Multi-Criteria Analysis demonstrate the highest adaptability, as their flexible frameworks can be readily customized by modifying criteria or preference structures to accommodate diverse evaluation requirements.

The Strategic Compatibility Framework shows medium adaptability, as it is best suited for high-stakes strategic partnerships and less applicable to operational or transactional relationships. Similarly, Profiling and Gap Analysis is most valuable for capability-focused partnerships. While both can be adapted, their inherent focus makes them less universally applicable. The KPI-based Decision Matrix also has medium adaptability, as its effectiveness is limited by the availability and relevance of standardized performance indicators across different industries, benefiting most from contexts where such metrics are common and reliable [6].

2.2.3 Synthesis Matrix

Methodology Complexity Resources Scalability Transparency Adaptability Best Use Case Weighted Scoring Matrix Low-Medium Low-Medium High Very High High Structured procurement with multiple criteria Strategic Compatibility Framework High High Low Medium Medium High-stakes strategic alliances and joint ventures Multi-Criteria Analysis (MCA) Medium-High Medium-High Medium High High Complex decisions with conflicting stakeholder priorities Profiling & Gap Analysis Medium Medium-High Low-Medium High Medium-High Capability-based partnerships and technology alliances KPI-based Decision Matrix Low-Medium Medium Very High Very High Medium Performance-critical operational partnerships Hybrid ML-based Matchmaking (with XAI) High High Very High Medium-High (depends on dataset quality) Very High Digital business ecosystems, automated partner/service recommendation with explainability

2.2.4 A context-dependent framework for Methodology Selection

The optimal choice among the presented methodologies is not absolute but is dependent on several key factors related to the decision context, organizational capabilities, and strategic objectives. Based on established principles in decision analysis [6], a context-dependent framework can be proposed to guide the selection process. Decision Urgency and Timeline: Time-sensitive decisions favor efficient methodologies like Weighted Scoring Matrix or KPI-based Decision Matrix. Strategic partnerships requiring deep analysis justify the time investment in Strategic Compatibility Framework or comprehensive Profiling and Gap Analysis.

Partnership Strategic Importance: High-stakes strategic partnerships warrant sophisticated analysis using Strategic Compatibility Framework or Multi-Criteria Analysis. Operational partnerships may be adequately served by Weighted Scoring Matrix or KPI-based approaches.

Stakeholder Complexity: Multi-stakeholder decisions with potentially conflicting priorities benefit from Multi-Criteria Analysis's structured approach to tructured process for defining preferences and trade-off management. Simpler stakeholder structures may be adequately served by more straightforward approaches.

Available Information and Data: Data-rich environments enable KPI-based analysis, while information-scarce contexts may require profiling-based approaches that generate necessary insights through structured assessment.

Organizational Analytical Capabilities: Organizations with strong analytical capabilities can leverage sophisticated methodologies like MCA or comprehensive strategic analysis. Less analytically sophisticated organizations may prefer transparent, straightforward approaches.

Industry Context: Industries with standardized performance metrics favor KPI-based approaches. Knowledge-intensive sectors may benefit from profiling and gap analysis. Regulated industries may require transparent, auditable methodologies.

2.2.5 Hybrid Approaches

While presented separately, these methodologies need not be used in isolation. Hybrid approaches that combine elements from multiple methodologies can address the limitations of individual methods while leveraging their respective strengths. Common hybrid strategies include sequential filtering, parallel assessment, and integrated frameworks .

Sequential Filtering employs different methodologies at different stages of the partner selection process. For instance, KPI-based screening might identify a long list of qualified partners, followed by Strategic Compatibility Framework analysis for shortlisted candidates. This approach balances efficiency with depth, enabling broad initial search followed by detailed evaluation of promising candidates. A typical sequential approach might involve:

1. Initial Screening: KPI-based filtering to identify partners meeting minimum performance thresholds. 2. Detailed Assessment: Profiling and Gap Analysis to understand complementarity potential. 3. Strategic Evaluation: Strategic Compatibility Framework for final candidate selection. 4. Implementation Planning: Multi-Criteria Analysis to optimize partnership structure and governance.

Parallel Assessment applies multiple methodologies simultaneously, using convergent findings to increase confidence and divergent findings to surface important considerations. An organization might conduct both Weighted Scoring Matrix and Profiling and Gap Analysis exercises, using agreement between methods to validate decisions and disagreement to prompt deeper investigation .

Integrated Frameworks combine elements from different methodologies into unified approaches. For example, Multi-Criteria Analysis might incorporate strategic compatibility as one criterion category while using KPIs as performance measures for other criteria. This integration requires careful design to ensure methodological coherence but can produce comprehensive assessment frameworks tailored to specific organizational needs .

2.3 Conclusions

2.3.1 Synthesis of Key Insights

The exploration of qualitative methodologies for B2B matchmaking reveals a rich landscape of approaches, each offering unique perspectives on the fundamental challenge of identifying and selecting compatible business partners. The diversity of methodologies reflects the multifaceted nature of B2B relationships and the varying contexts in which partnership decisions occur. No single methodology emerges as universally superior; rather, each provides valuable tools suited to particular circumstances and objectives.

A critical insight emerging from this analysis is that successful B2B matchmaking requires balancing analytical rigor with practical implementability. While sophisticated methodologies like Multi-Criteria Analysis and Strategic Compatibility Framework offer comprehensive treatment of complex partnership dynamics, simpler approaches like the Weighted Scoring Matrix often prove more effective in practice due to their transparency and ease of implementation. Organizations must honestly assess their analytical capabilities and stakeholder requirements when selecting methodologies.

The importance of context in methodology selection cannot be overstated. Factors such as industry dynamics, partnership objectives, available information, time constraints, and organizational capabilities all influence which approach will prove most effective. The comparative analysis reveals that methodologies vary significantly in their resource requirements and scalability, making some more suitable for broad partner screening while others excel at deep strategic assessment.

The complementary nature of different methodologies suggests that hybrid approaches may offer the most promise for complex partnership decisions. By combining the strengths of different methodologies while mitigating their individual limitations, hybrid approaches can provide both efficiency and depth, objectivity and insight, breadth and focus as circumstances require.

2.3.2 Implications for Practice

Beyond the selection of a specific methodology, the preceding analysis yields several cross-cutting strategic implications for professionals seeking to enhance their matchmaking processes and outcomes [10].

Methodology Selection Should Be Strategic: The selection of matchmaking methodology should be deliberate and informed, based on careful consideration of decision context and requirements. Organizations should resist defaulting to familiar approaches without evaluating their appropriateness for specific partnership challenges. The comparative framework presented in this chapter provides a systematic approach to methodology selection based on organizational capabilities and decision requirements.

Investment in Process Design Produces Long-Term Value: Regardless of methodology selected, careful attention to process design significantly influences outcomes. This includes stakeholder engagement in criteria development, clear definition of evaluation procedures, training of evaluators, and establishment of quality assurance mechanisms. Organizations that invest in robust process design achieve more consistent and defensible partnership decisions.

Capability Development Enables Sophistication: Organizations seeking to improve their partnership selection capabilities should consider developing analytical competencies that enable more sophisticated methodologies. This might involve training in strategic analysis, methods for assessing stakeholder preferences, or data analytics. The investment in capabilities enables organizations to tackle more complex partnership challenges and achieve better outcomes.

Transparency Facilitates Organizational Learning: Methodologies that provide transparent, auditable decision processes enable organizational learning and continuous improvement. By documenting decision rationale and tracking partnership outcomes, organizations can refine their selection criteria and improve future decisions. This learning orientation transforms partnership selection from isolated decisions to systematic capability development.

Integration with Broader Processes Enhances Value: Partnership selection methodologies achieve greatest impact when integrated with broader strategic planning, capability development, and performance management processes. This integration ensures that partnership decisions align with organizational strategy and that partnership outcomes inform future strategic choices.

2.3.3 Future Directions and Transition to Application

The five methodologies analyzed in this chapter provide a robust set of frameworks for judgment-based B2B matchmaking. They are thorough, grounded in established theory, and effective in contexts where human expertise is essential.

A plausible trajectory for the evolution of matchmaking appears to be the integration of automation and Artificial Intelligence to accelerate and scale the partner selection process. This path, however, is not without significant challenges. While AI models can process vast amounts of data and identify patterns far beyond human capability, they often function as opaque "black boxes." In the high-stakes context of B2B partnerships, where decisions can have significant financial and strategic consequences, relying on a recommendation from an inscrutable algorithm can be problematic. Trust, accountability, and the need to justify strategic choices to stakeholders remain important considerations.

This tension suggests a crucial direction for future work: the challenge may no longer be simply about increasing the predictive accuracy of matchmaking models, but about ensuring their explainability. Consequently, the value of future methodologies could be judged not merely on their accuracy, but on their capacity for transparent and demonstrably fair operation.

Before such systems can be built, however, a foundational question arises regarding what truly constitutes a meaningful and trustworthy explanation for business professionals. Investigating these human-centric aspects of explainability therefore emerges as a critical research priority. This challenge, central to the present thesis, will be explored in subsequent chapters through an empirical investigation with industry experts.

3

Methodology

3.1 Introduction to the Problem and Methodological Approach

The evolution of B2B matchmaking toward data-driven systems has surfaced a central methodological problem: the opacity of complex AI models. In enterprise settings, opaque decision pipelines hinder accountability and can erode stakeholder trust. Explainable AI (XAI) provides the methodological basis to make model behaviour inspectable and to preserve transparency where expert judgment remains decisive. At the same time, handling sensitive corporate information raises material risks for privacy and security that any methodological design must control ex-ante.

The goal in this thesis is not to replace human expertise, but to augment it with tools that are transparent, auditable, and privacy-preserving. To this end, I develop a decision-support tool for B2B matchmaking experts that is engineered around two methodological pillars.

First, a strict Privacy-by-Design stance is operationalized through the exclusive use of a synthetic dataset for model development and testing. This choice reduces the exposure surface for real corporate data, limiting the points where a breach could reveal identifiable information while still enabling systematic experimentation. Second, the decision support itself combines efficient retrieval with post-hoc explanation. A dual-encoder (“two-tower”’) deep learning architecture is used to screen large partner spaces and to produce a shortlist of high-potential candidates, a task that is otherwise time-consuming and error-prone. On top of this retrieval stage, XAI techniques are employed to generate case-level explanations that make the salient drivers of each recommendation explicit for the domain expert. This design is motivated both by the need to foster user trust and by regulatory principles that require meaningful information about the logic involved in automated decision-making (GDPR information rights and Article 22safeguards).[1]

Methodologically, the matchmaking task is framed as an information-retrieval problem: given a query entity (a beneficiary with a need), the system must identify the most relevant candidates (suppliers) in a large corpus. The deep learning approach is chosen for its ability to learn semantic representations of organizations and needs, overcoming the brittleness of keyword-based methods and enabling more faithful similarity assessments.

The remainder of this chapter details the data generation and validation strategy for the synthetic corpus, the model architecture and training protocol, and the explanation layer, including the evaluation metrics used to assess both predictive performance and interpretability.

3.2 The Dataset

3.2.1 Generation of the Synthetic Dataset

Access to real company data for innovation matchmaking is limited by confidentiality and contractual constraints. To enable empirical work without processing sensitive information, the study relies on a synthetic corpus.

The dataset reproduces the structure commonly found in innovation platforms and public calls. It contains two base tables: one that describes organizations expressing a technological need (beneficiaries) and one that profiles potential providers. Fields represent domain-relevant dimensions rather than any real record: sector or industry, a set of technical skills, indicative technology readiness, indicative project costs, and basic organizational attributes.

Supervision is provided by a label table generated by rules that formalize what “compatibility” means for this task. Positive pairs are created only when three conditions hold at the same time:

1. **Sector–skill alignment.** The beneficiaryś declared need must be present among the supplierś technical skills. The generator uses a curated mapping that links sectors to families of skills (for example, IoT, AI, Cybersecurity).

[1] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation).* Official Journal of the European Union. Articles 13–15 and 22. 2016.

2. **Technology readiness fit.** The supplierś average Technology Readiness Level must meet or exceed the beneficiaryś minimum requirement.

3. **Budget coherence.** The supplierś average project cost must not exceed the beneficiaryś maximum budget.

Pairs that do not satisfy all requirements become negatives. This rule-based labeler makes the provenance of each label auditable and allows reviewers to trace exactly why a pair is considered a good match.

Because the space of all possible pairs is large, non-matches outnumber matches. To keep training informative, the dataset includes randomly sampled negatives and, where appropriate, *near-miss* negatives that violate only one rule (for example, a small budget gap). These examples help the model learn useful decision boundaries rather than trivial distinctions. Taken together, the synthetic corpus and its rule-based labels provide a transparent, auditable, and privacy-preserving foundation for the study. The dataset encodes an explicit, reproducible notion of compatibility independent of proprietary records.

3.2.2 Exploratory Data Analysis: Understanding the Simulated Ecosystem

An Exploratory Data Analysis (EDA) was conducted to assess face validity—whether the synthetic corpus behaves like a plausible innovation ecosystem—and to extract practical cues for preprocessing. On the beneficiary side, the population comprises 1,000 unique entities. A sizable share operates in Manufacturing (16.1%), which aligns with the expectation that legacy sectors show sustained demand for technological upgrading.

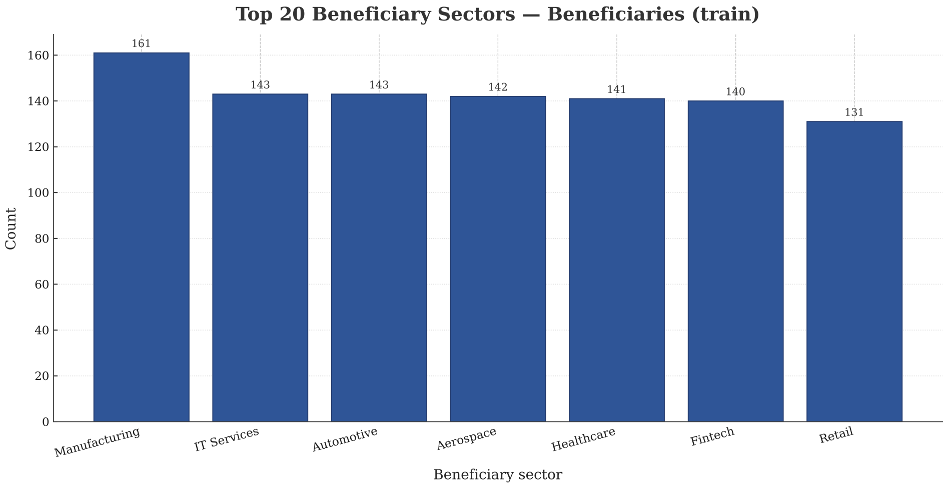

Figure 3.1: Distribution of beneficiary sectors in the training set, showing a significant concentration in Manufacturing.

Declared needs are not uniformly spread but concentrate on Big Data (22.3%), consistent with organization-wide efforts toward data-driven efficiency.

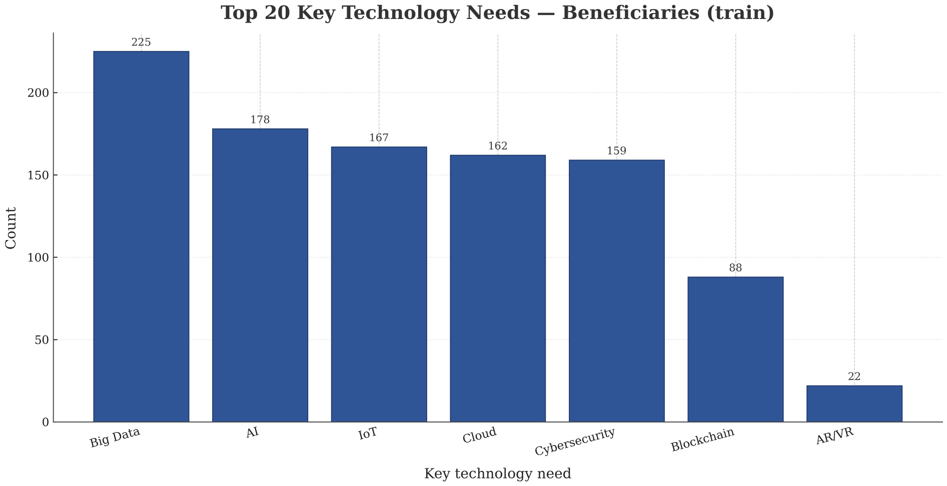

Figure 3.2: Distribution of key technological needs among beneficiaries, highlighting the prevalence of Big Data.

Supplier profiles span multiple industries and skill sets.

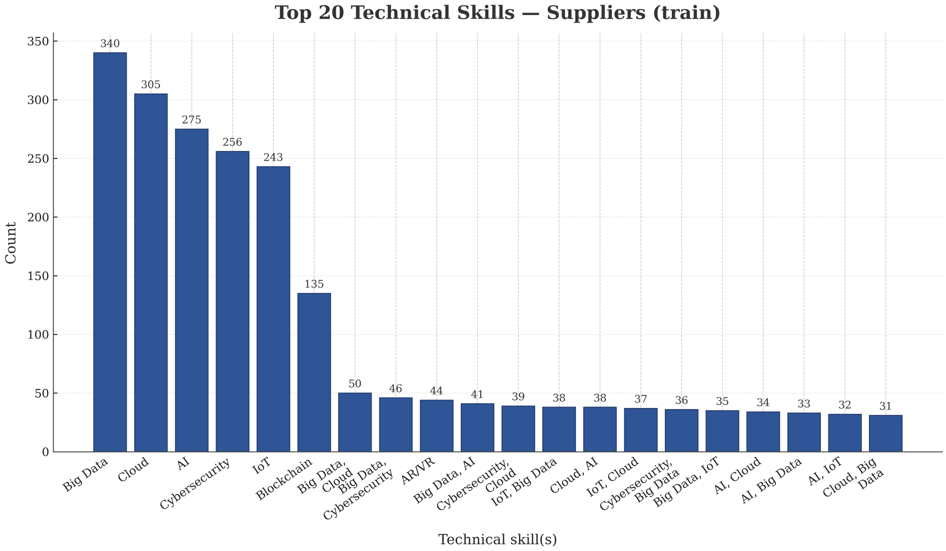

Figure 3.3: Distribution of technical competencies across the supplier population.

Rather than reporting exhaustive counts, the EDA focuses on structural properties that matter for modelling: the presence of multi-label skill sets per supplier, overlaps between sector and skill vocabularies, and the degree to which beneficiary needs are covered by the available competencies. Label consistency was also verified by checking that every positive pair satisfies the generation rules without contradictions on TRL or budget, and by ensuring that text fields meet minimum length and tokenize correctly so as not to inflate performance through artefacts. Overall, the corpus appears varied enough to train a retrieval model and sufficiently regular to support stable learning, which in turn motivates the preprocessing choices described next.

3.2.3 Data Preprocessing and Splitting: Translating Business Concepts into a Machine-Readable Form

Data preparation proceeds as a careful translation of heterogeneous business descriptors into model-ready representations. Numerical attributes such as employee counts and budgets are standardized so that differences in scale do not dominate optimization; where appropriate, mild clipping of extreme values stabilizes gradients while preserving rank information. Categorical and textual signals are encoded through embeddings rather than sparse one-hot vectors. Multi-label skill sets are represented as learned vectors aggregated with simple pooling operators, while short free-text descriptions undergo minimal normalization to preserve domain terminology before tokenization and embedding. This pipeline allows the model to learn semantic proximity among technologies and domains—for instance, positioning Cloud and AI closer to one another than to Logistics—instead of treating categories as independent atoms.

To obtain a clean estimate of generalization, the corpus is partitioned into train (80%) and test (20%), with 10% of the training portion held out for validation. Splits are *entity-disjoint*, so that the same beneficiary or supplier never appears in more than one split; where feasible, stratification by sector maintains comparable distributions across folds. This protocol prevents leakage through overlapping descriptions or attributes and focuses evaluation on unseen entities. Because non-matches vastly outnumber matches, training employs negative sampling at a ratio of 1 positive : 4 negatives. Alongside randomly drawn non-matches, the sampler includes *near-miss* negatives that violate only a single rule (e.g., a slight budget excess), which sharpens the decision boundary and makes the retrieval task more informative. Sampling affects training only; validation and test preserve their natural prevalence (or the fixed candidate sets specified by the evaluation protocol) so that reported metrics are not biased by resampling choices. The resulting tensors feed the dual-encoder architecture introduced in the next section, with categorical and numerical representations learned end-to-end and text represented via a frozen sentence encoder.[2]

3.3 Model Architecture and Training: Engineering a System for Semantic Compatibility

The task is large-scale candidate retrieval: for each beneficiary, the system must surface a short, relevant list of suppliers from a broad pool. A Two-Tower design is used so that a beneficiary (query) and each supplier (candidate) are represented by separate networks that map into the same vector space. Compatible pairs end up close to one another, which enables fast top-k retrieval. This design follows established practice in large-scale recommendation, where decoupling query and candidate encoders supports efficient retrieval at inference time.[3]

[2] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[3] Paul Covington, Jay Adams, and Emre Sargin. "Deep Neural Networks for YouTube Recommendations". In: *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. Boston, MA, USA: ACM, 2016. DOI: 10.1145/2959100.2959190; Google Developers. *Two-Tower Retrieval*

The separation of towers also serves an engineering goal: supplier vectors can be computed in advance and stored in a vector index, while at query time only the beneficiary vector is computed and nearest neighbours are retrieved in real time.[4] Each tower integrates different kinds of information and fuses them into a single representation. Numerical fields (for example, budgets or firm size) are standardized and passed through simple dense layers to avoid scale dominance. Short textual fields are converted into vectors with a pre-trained Sentence-BERT encoder kept frozen, a choice that keeps the pipeline reproducible and avoids overfitting on a narrow, synthetic corpus.[2] Categorical fields (such as sector or region) are mapped via lookups to trainable embeddings, with embedding sizes scaled to vocabulary size so that capacity matches feature complexity. The fused representation is projected to a compact vector with standard regularization (L2) and Batch Normalization.[5] Similarity between beneficiary and supplier vectors is computed with a normalized dot product (cosine-like) measure, which focuses learning on directional alignment rather than raw magnitude.[6]

The model was improved iteratively to obtain the best configuration possible given project and hardware constraints. A point-wise Binary Cross-Entropy objective on individually sampled pairs was favored over list-wise losses because it is simple to implement, stable under limited compute, and naturally compatible with controlled negative sampling.

For text, early bag-of-words/TF-IDF baselines proved brittle to vocabulary shifts, whereas a frozen sentence encoder delivered more stable behavior on short descriptions. Cross-encoders were considered as a re-ranking stage on the shortlist, but their runtime cost and integration complexity placed them outside the scope of this prototype. [6]

Negative design is integral to the learning setup. Because non-matches vastly outnumber matches, negative sampling at a 1:4 ratio is applied in training only, combining random negatives with near-miss negatives that violate a single rule (for example, a slight budget excess) to sharpen the decision boundary. Validation and test preserve natural prevalence and follow the entity-disjoint protocol introduced earlier so that generalization

for Recommendation. Accessed 2025-09-14. 2022. URL: https://developers.google.com/machine-learning/recommendation/two-tower.

[4] Google Developers, see n. 3.

[5] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. 2015.

[6] Covington, Adams, and Sargin, see n. 3.

is measured on unseen entities and leakage is avoided. Regularization combines Dropout and EarlyStopping monitored on validation. All preprocessing components are fit on the training split and then frozen for validation, test, and inference to ensure consistency and auditability. Training runs on local hardware, which limited exhaustive hyper-parameter search; to preserve comparability across experiments, the configuration is kept fixed, with advanced tuning deferred to the limitations section.

3.4 Evaluation Methodologies: Measuring Effectiveness and Ensuring Transparency

3.4.1 Measuring Effectiveness

Effectiveness is assessed with metrics aligned to the workflow of reviewing recommended partners:

- **ROC–AUC** condenses the model's ability to separate positive from negative pairs across all possible decision thresholds. Values close to 1 indicate strong separability, while 0.5 corresponds to chance.[7]

- **Precision** consists in the share of predicted matches that are actually correct. High precision means the system rarely surfaces weak or irrelevant pairs; higher thresholds typically increase precision at the cost of recall.

- **Recall**, instead, is the share of all true matches that the system successfully retrieves. High recall means fewer missed opportunities; lower thresholds typically increase recall at the cost of precision.

- **F1** combines precision and recall into a single number (harmonic mean), penalizing configurations that improve one metric only by sacrificing the other.

Thresholds are calibrated on the validation set to reflect the desired balance between reviewer workload (favoring precision) and opportunity capture (favoring recall). All metrics are computed on held-out, entity-disjoint splits with natural label prevalence and without carrying over any training-time resampling.

[7] Tom Fawcett. "An Introduction to ROC Analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

At inference time, recommendations are served via an ANN index to meet latency and scale requirements. Because approximation can hide part of the model's signal, evaluation distinguishes two settings. In the *full-recall* setting, similarities are computed against the entire candidate pool without using the index; this captures the model's upper-bound discriminative behavior. In the *partial-recall* setting, the ANN index is used exactly as in production and metrics are computed on what the index returns. To keep this measurement honest, the protocol reports the index's recall (the fraction of true nearest neighbours that the index successfully retrieves at the chosen operating parameters) together with precision, recall, and F1. Index hyperparameters are selected on the validation set and then frozen for testing.[8]

3.4.2 Explainability Framework

Explainability is central to the methodology. In the context of B2B innovation matchmaking, where strategic decisions carry significant financial and operational implications, the "black box" nature of deep learning models can erode trust and hinder expert adoption. Therefore, XAI was not merely an optional enhancement but a fundamental requirement for this project, aiming to provide human experts with transparent insights into why specific recommendations were made, thereby fostering trust, enabling critical validation, and facilitating informed decision-making. This section details the practical application of XAI techniques within the prototype, focusing on LIME for local explanations, the challenges with SHAP, and the exploration of Integrated Gradients as an alternative.

LIME: Local Interpretability for Individual Recommendations. To shed light on the modelś decisions at the level of individual predictions, the LIME (Local Interpretable Model-agnostic Explanations) technique was implemented.[8] LIMEś core logic is to understand a complex modelś prediction by probing it with slightly modified versions of the input. For any given prediction, LIME generates a new dataset by creating small "perturbations" around the original input instance. It then obtains predictions from the black-box model for these perturbed samples. By observing how the modelś output changes with these small variations, LIME trains a simple, interpretable model (like a

[8] Jeff Johnson, Matthijs Douze, and Hervé Jégou. "Billion-Scale Similarity Search with GPUs". In: *IEEE Transactions on Big Data*. Vol. 7. 3. 2019, pp. 535–547. DOI: 10.1109/TBDATA.2019.2921572.

linear regression) on this local neighborhood. The coefficients of this local model then reveal which input features had the greatest positive or negative influence on that specific prediction. LIMEś model-agnostic nature allows it to be applied to any machine learning model, providing local fidelity by accurately reflecting the modelś behavior in the vicinity of the explained instance. The implementation of LIME presented several practical challenges, as documented during the development process. These included: - Data Type Mismatches: Ensuring that categorical features, including entity IDs, were correctly converted to integer codes for LIMEś `LimeTabularExplainer`, while simultaneously ensuring the modelś predict function could correctly map these back to original string IDs for embedding lookups. - Preprocessor Inconsistencies: Resolving issues where LIMEś internal scaling mechanisms conflicted with the modelś expected input formats, requiring explicit type conversions and careful management of pre-calculated embeddings. These challenges underscored the practical complexities of integrating model-agnostic XAI techniques with deep learning pipelines, particularly when handling heterogeneous data.

To avoid repetitive outputs in a two-tower setup, raw entity identifiers are excluded from the explainer inputs; the explainer mirrors the model's preprocessing so that perturbations remain in-distribution. Known sensitivities—how the neighbourhood is sampled and potential instability—are mitigated by fixing seeds, using consistent sample sizes, and constraining values to valid ranges. When text is represented as a single embedding vector, the explanation reports the contribution of the text channel as a whole; a token-level view can be produced on demand using a text-specific explainer with the same frozen encoder.[3]

SHAP: Shapley Additive Explanations. While SHAP was initially considered as the primary XAI technique due to its strong theoretical guarantees rooted in cooperative game theory, its full implementation proved to be a significant technical and computational hurdle for this project.[9] SHAP values quantify the average marginal contribution of each feature value to the prediction across all possible coalitions of features. This provides a unified and theoretically sound measure of feature importance, offering several key advantages: - Theoretical Guarantees: SHAP values are the only method with a solid theoretical foundation, ensuring properties like local accuracy, consistency, and missingness. - Global Interpretability: By aggregating local SHAP explanations, it is possible to derive global insights into which features are most important overall for the modelś predictions. - Consistency: If a model changes such that a feature has a greater impact, its SHAP value will reflect this.

However, applying SHAP algorithms, particularly `KernelSHAP` or `DeepSHAP`, to the Two-Tower architecture with its heterogeneous data (numerical, categorical, and BERT embeddings) presented profound technical and computational challenges that rendered its comprehensive application practically unfeasible within the projectś scope. The Two-Tower modelś structure, where beneficiary and supplier features are processed by separate neural networks and only interact at the final cosine similarity layer, fundamentally complicates the direct application of SHAPś more efficient variants. `DeepSHAP`, for instance, is designed for single-input, end-to-end neural networks and struggles to attribute contributions across two distinct, independently processed input streams that converge only at the very end. `KernelSHAP`, while model-agnostic, is notoriously slow, requiring a vast number of model evaluations for each explanation. This burden is compounded by the high dimensionality of the combined feature space and the complexity of explaining features that are transformed into dense embeddings early in the pipeline. These technical barriers, inherent to the modelś architecture and data complexity, were the primary reason for not fully implementing SHAP.

Integrated Gradients (IG): Gradient-Based Attributions on a Tailored Model. IG is a feature attribution method that calculates the importance of each input feature to a modelś prediction.[17] Its core logic involves integrating the gradients of the modelś output with respect to its inputs, along a linear path from a defined baseline (a neutral input, typically a zero or null vector) to the actual input instance being explained. Given the technical challenges encountered with applying SHAP to the full Two-Tower model, Integrated Gradients (IG) was explored as an alternative approach for feature attribution. This exploration was specifically conducted on a version of the Two-Tower model tailored for gradient-based analysis, focusing exclusively on numerical features and pre-calculated text embeddings, while deliberately excluding categorical features. This strategic simplification allowed for a focused investigation into the attributions of continuous and textual features, enabling compatibility with gradient-based XAI methods in the local development environment, particularly given the complexities introduced by

`Embedding` and `BatchNormalization` layers in the full model. This method satisfies important axioms, such as completeness, which ensures that the sum of the attributions across all features equals the difference between the modelś prediction for the input and its prediction for the baseline.

4

Results and Discussion

4.1 Validation in the Cyber 4.0 Context

To assess real-world applicability, validation was designed to mirror the operational and governance conditions of the proposed project. Cyber 4.0 is Italy's National Highly Specialized Competence Center for Cybersecurity—one of eight competence centers established and co-funded by the "Ministero delle imprese e del Made in Italy"—and is recognised as a national technology-transfer hub and implementing party for the PNRR. Its mandate covers training, guidance, and applied innovation, delivered through a public–private partnership that includes universities, research bodies, large firms and SMEs. In this setting, B2B matchmaking between technology suppliers and beneficiaries is a core workflow and must satisfy requirements of security, trustworthiness, and privacy-by-design.

Accordingly, a single-expert validation session was conducted using Cyber 4.0 as the reference context. The session combined: (i) a briefing on the analyzed matchmaking technique and an explanation of the prototype; (ii) a structured post-session questionnaire on trust, transparency and data protection; and (iii) an open discussion on adoption conditions.The participant came from an institutional context focused on technology transfer in the B2B domain, where the effectiveness of initiatives is typically measured by the number of partnerships successfully concluded. The design of the validation thus privileged depth on governance-critical aspects over broad sampling. In practice, the expert framed success in terms of partnership agreements reached, and their reaction to the AI matchmaker concept was highly favorable, assigning it the maximum score on the questionnaire's rating scale. This feedback indicates a clear requirement: the prototype must help convert qualified candidates into signed agreements while making that contribution visible. The phases that currently absorb the most effort indicate where the tool should act: a guided intake to surface real—often unspoken—needs; an assisted search constrained by skills and TRL; a compatibility view highlighting budget fit, skills overlap, and TRL gaps; and a lightweight monitoring panel for active matches.

The expert emphasized costs and budget, technical fit and innovation, partner reliability and reputation, and contractual and intellectual-property aspects, while strategic and cultural alignment was considered important though negotiable. These priorities also indicate how explanations should work. Rather than generic feature weights, each recommendation should deliver a short strategic-compatibility summary expressed in those criteria, cite comparable past cases that justify the suggestion, and expose a calibrated confidence score. Deal-breakers such as poor reputation and the difficulty of accessing honest feedback or true technology maturity imply adding reputation signals (e.g., ratings or scores), links to TRL evidence, and structured prompts that capture needs and post-project feedback.

The prototype is a solid base to address the needs surfaced in the expert consultation, primarily by tackling the problem of "black box" AI in high-stakes decisions. While its development on a synthetic corpus provides a strong privacy-by-design foundation, the prototype's core value proposition in a business context is its commitment to transparency. By providing local, actionable explanations for each match—expressed in the same language used by practitioners (budget fit, skills overlap, TRL gaps)—it makes the AI's reasoning clear. This explainability is crucial for keeping the human expert in the loop, allowing them to apply their own contextual knowledge and maintain ultimate control over the final decision. This *human-in-the-loop* architecture is the primary safeguard that ensures alignment with Article 22 of the GDPR. The article grants individuals the right not to be subject to a decision based *solely* on automated processing which produces legal or similarly significant effects. Because the prototype is designed as a decision-support tool—where the AI recommends but the human expert makes the final determination—no decision is ever made "solely" by the automation. The transparency provided by the XAI methods further reinforces this, empowering the expert to critically evaluate and, if necessary, override the model's suggestion, thus maintaining meaningful human control at all times. Furthermore, the prototype addresses practical adoption barriers with a light, self-contained setup based on simple CSV/Excel ingestion, targeting the most time-consuming phases of the matchmaking workflow.

Beyond validating the current approach, the expert consultation also provided a rich set of suggestions to meet the need of the company and the specific context. A central suggestion was the creation of a post-match feedback loop. This would involve developing a system to capture the real-world outcomes of partnerships, allowing the model to learn continuously from which matches were truly successful. Such a mechanism would progressively enhance the model's accuracy by grounding its predictions in tangible business results. The discussion also turned to the evolution of the explainability system, envisioning a move beyond case-by-case justifications towards the extraction of more generalizable strategic rules. This would transform the explanations from a simple validation tool into a source of broader business intelligence, revealing deeper insights into the factors that consistently drive successful partnerships. Finally, to address the performance limitations of the initial prototype, the exploration of alternative model architectures was recommended as a path to better capture the intricate patterns within the data. These suggestions form the basis for the future work discussed in the final chapter.

The expert distinguished two operative postures for Cyber 4.0: acting as a neutral third-party broker between external organisations (institutional mandate) and acting as an active party in the match (commercial posture). The governance implications differ. The former calls for neutrality controls, audit trails, and explanations oriented to bilateral accountability; the latter requires clear declarations of interest, a visible "mode" indicator in the UI, and performance tracking tied to business-development goals. We reflect these distinctions in the pilot plan (separate evaluation dashboards and audit logs per posture) and in the roadmap (reputation signals, TRL evidence links, case-based justifications, and a calibrated confidence score).

These findings support a human-in-the-loop flow. Explanations should be brief, checkable, and linked to precedent, and the UI should expose a confidence score by default. The expert's feedback also confirmed that using synthetic data for development and piloting is an essential approach to protect confidentiality. Finally, features should exclude identifiers and centre on signals that reflect real decision criteria (budget fit, skills match, TRL). As a single-informant exercise, these results are formative and do not claim statistical generalisation; they align the prototype with practitioner priorities and define a concrete pilot-evaluation plan.

4.2 Quantitative Results and Model Performance

Quantitative evaluation aligns the prototype with its decision-support objective: recommending beneficiary–supplier matches with evidence that domain experts can review. The model adopts a two-tower retrieval architecture that combines sentence embeddings for text with categorical and numerical signals. Out-of-sample performance is measured on a held-out test split with strict leakage controls—entity-disjoint splits, negative sampling 1:4 confined to training, and identifiers excluded from features. The evaluation uses the curated synthetic dataset with rule-based labels adopted in the project.

Within a test distribution where the positive class (match) had a prevalence of 33%, the model performs well. With a standard decision threshold of 0.5, overall accuracy is **0.903**. For the positive class, **precision = 0.98**—recommended pairs are almost always correct—while **recall = 0.72**, meaning the model retrieves a large share of genuine opportunities. Overall discrimination is consistent with these figures, with **ROC–AUC = 0.86**.

Table 4.1: Test performance (held-out test)

Metric	Value	Scope
Accuracy	0.903	Overall
ROC–AUC	0.86	Overall
Precision	0.98	Class 1 (match)
Recall	0.72	Class 1 (match)

These values depict a *conservative recommender*: high precision with solid recall yields few false positives and good coverage of true opportunities. This profile suits a human-in-the-loop setting, where experts spend less time triaging noise and can review near-miss negatives with contextual knowledge.

Two considerations frame interpretation. First, labels in the synthetic corpus encode a rule-based specification of compatibility; strong results show that the model has learned that specification under controlled conditions. Second, synthetic data are typically more uniform and less noisy than real-world data, so performance is not expected to transfer unchanged to operational contexts. Within these boundaries, the evidence supports the prototype as a *sound starting point* for validation in realistic settings.

4.3 Explainability and Transparency Analysis (XAI)

4.3.1 Introduction to XAI Methods

To ensure transparency and build trust in the model's recommendations, two Explainable AI (XAI) techniques were employed: LIME (Local Interpretable Model-agnostic Explanations) and Integrated Gradients (IG). As detailed in Chapter 3, these methods provide local, instance-specific explanations for each prediction, turning the model from a black box into a transparent decision-support tool.

4.3.2 LIME: Output and Interpretation

LIME explains each prediction by fitting a simpler, interpretable linear model in the neighbourhood of the specific instance. The method produces a list of the most relevant features, each accompanied by a weight (importance score) that summarises its effect on the decision. In practice, positive weights push the model towards a "Match", negative weights towards a "No Match"; the larger the weight, the stronger the feature's contribution.

4.3.3 LIME Results and Analysis

The analysis confirms that the model's reasoning aligns with the expected business logic. For instance, in a "Match" prediction between Beneficiary A and Supplier B, LIME produced the following key feature contributions:

Table 4.2: Example LIME contributions for a "Match" instance

Feature	**LIME Weight**
Supplier's Main Sector	+0.25
Beneficiary's Key Tech Need	+0.20
TRL Compatibility	+0.15
Budget Compatibility	+0.10
Supplier's Legal Region	−0.05

Interpretation. These results translate into a clear business narrative: the pairing is favoured because the supplier's main sector is "Information Technology" and the beneficiary's key tech need is "Custom Software Development". This positive outlook is further supported by TRL and budget compatibility, while the supplier's non-local region acts as a minor negative factor.

4.3.4 Integrated Gradients (IG): Output and Interpretation

Integrated Gradients provides a faithful, feature-attribution explanation ideal for neural networks. It operates by accumulating a feature's gradients along a path from a neutral baseline (e.g., a zero-input) to the actual input. The resulting attribution scores are positive for features pushing towards a "Match" and negative for those pushing towards a "No Match". A key property of IG is *completeness*, meaning the attributions sum up to the model's final prediction score relative to the baseline.

4.3.5 IG Results and Interpretation

Applying IG to the same "Match" instance (predicted with a confidence of 0.85) yielded attribution scores that were conceptually similar to LIME's but with precise numerical values tied to the model's inner workings:

Table 4.3: Example IG attributions for the same instance

Feature	**IG Attribution**
Sector & technology alignment	+0.35
TRL compatibility	+0.25
Budget compatibility	+0.20
Other positives (e.g., work approach)	+0.10
Location (negative)	−0.05

Interpretation. The sum of these attributions ($0.35 + 0.25 + 0.20 + 0.10 - 0.05 = 0.85$) equals the model's prediction score, demonstrating completeness. The breakdown confirms that the most critical factors were, in order of importance: technological fit, TRL match, and budget alignment.

4.3.6 Conclusion

The XAI analysis demonstrates that the matchmaking model operates on a logical and understandable basis. Both LIME and IG confirm that the model prioritises key business criteria—technological alignment, TRL, and budget. This behaviour was consistent across multiple analysed instances, confirming that the model has learned a generalizable and sound business logic. The ability to generate clear, actionable explanations for each recommendation validates the model as a trustworthy and transparent tool suitable for a human-in-the-loop workflow. This provides a solid foundation for operational adoption, enabling experts to make faster, more confident decisions.

4.4 Limitations of the Study

This study was developed under clear technical and methodological boundaries that inform how results should be read relative to the thesis objective: a transparent, privacy-aware decision-support prototype suitable for enterprise settings.

4.4.1 Hardware Constraints and Experimental Scope

Development, training, and validation were conducted on CPU-only hardware. This constrained the experimental space, favored a lean two-tower architecture, and limited extensive hyperparameter sweeps. The choice was intentional—feasibility under realistic enterprise budgets—but it also narrows external validity compared with GPU-rich settings.

4.4.2 Data Strategy and Generalizability Risks

The dataset is synthetic by design to ensure controllability and privacy-by-construction. Generation rules mirrored real decision criteria (budget fit, skills overlap, technology readiness), yet they lack the noise, hidden confounders, and long-tail patterns of operational data. Consequently, strong in-sample performance may overstate robustness under heterogeneous and evolving inputs; distribution shift at deployment remains a credible risk.

4.4.3 Scope and Nuances of Explainability

Explainability was optimized for operations rather than theory-complete coverage. Local methods (LIME, Integrated Gradients) are sensitive to sampling and initialization, so repeated runs on the same instance can yield slightly different attributions. In this work, explanations are positioned as decision aids—useful for inspection and auditability—rather than ground truth about model causality. More global strategies (surrogates, rule extraction, large-scale SHAP) were assessed but deprioritized under the compute envelope to preserve responsiveness and usability.

4.4.4 Expert Validation and Key Findings

Within the Cyber 4.0 expert review, the prototype's matching logic aligned with practitioner criteria (economic viability, technical/innovation fit, partner reliability). The session surfaced these constraints explicitly and validated the research logic—data collected, criteria encoded, and explanation format—rather than the correctness of a specific model instance. Experts reported added value in guided intake, assisted search, a compatibility view that makes criteria checkable, and lightweight monitoring, all within a human-in-the-loop setup with concise rationales and confidence cues. In practice, the expert's feedback confirmed the value of a conservative but reliable recommender, observing that such a tool is fit for purpose in a human-in-the-loop workflow as it reduces triage effort. This supports the claim that, even under resource constraints and without personal data, it is possible to deliver useful and inspectable recommendations.

4.4.5 Synthesis of the Contribution

Therefore, the study's contribution should not be measured by the prototype's standalone performance, but by what it demonstrates about the process of building such a system under realistic constraints. Taken together, these limits advise caution in generalizing to production and clarify the next steps that will be detailed in Chapter 5: stress-testing the model under distribution shift and broadening the compute-aware explainability beyond local methods. The main outcome of this thesis is a prototype that demonstrates how a constrained setting—synthetic data, limited compute, and local explainability methods—can still deliver recommendations that are transparent, auditable, and meaningful in practice. Rather than a definitive solution, it should be seen as a proof of concept that highlights both the potential and the open challenges of explainable and privacy-aware decision support.

5

Conclusions

5.1 Overview

The research presented in this thesis originates from the critical challenge of deploying Artificial Intelligence within regulated enterprise contexts. The core objective is to resolve the inherent conflict between the high performance of complex “black box” models and the non-negotiable enterprise requirements for transparency, accountability, and compliance with data protection regulations like GDPR. The state-of-the-art review in Chapter 2 established this tension, outlining the evolution from interpretable systems to powerful but opaque deep learning models, and the subsequent rise of the Explainable AI (XAI) field to address this gap.

Informed by this context, this thesis argues that transparency, achieved through the practical application of XAI, is the primary mechanism for building user trust and ensuring meaningful human oversight over automated decisions. To explore this argument, a prototype matchmaking tool was developed, characterized by its “Privacy-by-Design” construction and its core integration of XAI techniques. The validation of this prototype yielded a nuanced but crucial outcome: rather than achieving perfect predictive accuracy, the project demonstrated the value of a holistic and responsible methodological framework, an approach whose value was confirmed by expert feedback which highlighted the importance of transparency and privacy in building user trust.

5.2 Answer to the Research Question

The central research question of this thesis asks: is it possible to implement an effective AI tool in an enterprise context while respecting regulatory constraints on privacy and transparency? This work provides an affirmative answer, demonstrating a practical possibility to be adopted. The argument rests on two distinct pillars: first, showing that regulatory privacy constraints can be met by adopting a development methodology that does not require processing sensitive data; and second, that the challenge of transparency can be overcome by integrating explainability directly into an expert's workflow, turning the AI into a trusted, auditable partner.

The first pillar, respecting privacy constraints, was addressed through a methodological choice that insulated the project from sensitive data. By successfully developing and validating the prototype on a synthetic dataset, this thesis demonstrates a viable path for enterprises to innovate with AI. It proves that building a high-performing model is possible without the significant legal and security risks of processing raw customer data. While this project used a synthetic dataset, the underlying principle could be implemented in a live enterprise context through other privacy-enhancing techniques, such as the robust encryption or anonymization of sensitive data fields.

The second pillar, achieving transparency, directly confronts the "black box" problem. As the results in Chapter 4 showed, the integration of XAI tools like LIME and Integrated Gradients rendered the model's decision-making process understandable. This transparency is not merely a technical feature; it is the mechanism that enables meaningful human oversight. By providing clear, actionable reasons for each recommendation, the system empowers the human expert to critically evaluate the AI's output, apply their own contextual knowledge, and ultimately maintain control. This directly addresses the spirit of regulations like GDPR's Article 22 by ensuring that the final decision is never based solely on automated processing.

This practical success leads to the second core debate raised in Chapter 2: the argument, most forcefully made by Cynthia Rudin, that we should prioritize inherently interpretable models over applying post-hoc explanations to "black boxes." While the appeal of a fully transparent model is strong, this thesis provides a crucial real-world counterpoint.

The matchmaking task at the heart of this project relies on understanding the rich, unstructured text of company descriptions, technological needs, and strategic objectives. As argued in the literature, neural network architectures are essential for this task, as they can capture the deep semantic nuances of language in a way that simpler models cannot. The "black box" was, therefore, a necessary choice to achieve a useful result. It allows the system to grasp the meaning of nuanced, written text, enabling experts to contribute valuable insights that cannot be constrained to a simple category or number.

Consequently, this work demonstrates that a complex model paired with a robust XAI layer is a powerful and pragmatic compromise. The prototype's high precision and the positive expert validation show that this "human-in-the-loop" approach is highly effective. It achieves the performance of a complex model while providing the necessary transparency to build trust and ensure accountability, offering a practical solution where purely interpretable models may fall short.

5.3 Contributions and Implications

The primary contribution of this thesis is intended not as a single algorithm, but as a **comprehensive** and replicable methodology for creating and validating AI systems within sensitive enterprise domains. This approach seeks to provide a practical guide for balancing the competing demands of model performance, traceability, and regulatory compliance. It suggests a concrete pathway for innovation that does not require compromising on core tenets like privacy and transparency, aiming to address the challenges that often stall such data-driven projects.

For firms and practitioners, the value of this paradigm *could be* immediate and tangible. First, it *can offer* a clear plan for conducting meaningful AI experimentation in regulated industries without processing raw personal or confidential data, potentially lowering the legal, security, and financial barriers to entry. Second, it *helps to move* Explainable AI from an abstract concept toward an applied instrument for the front line. The work *explores* how to integrate explanations that business experts can act upon, with the goal of fostering trust and improving the quality of human-in-the-loop decisions. Finally, the high-precision nature of the prototype *indicates that* this responsible model can still yield significant business value by reducing the manual effort required to find high-quality opportunities.

For policy makers and regulatory bodies, this research *may be viewed as* a valuable case study in responsible innovation. It *aims to illustrate* how systems can be engineered to be "compliant-by-design," where foundational principles like *data minimization* and *transparency* are woven into the project's structure from its inception. This *can provide* a positive example of how the goals of technological advancement and the aims of robust data protection might be aligned, *suggesting* a counter-narrative to the often-assumed conflict between the two.

5.4 Future Vision

5.4.1 A Role-Based, Context-Aware System

The future vision for this work is guided not only by technological possibility but also by the nuanced operational realities of an enterprise, as highlighted by the expert feedback. The suggestion to consider two distinct matchmaking roles—a neutral "Institutional Broker" and an active "Commercial Partner"—provides a powerful framework for a truly "out of the box" evolution of the prototype. The next generation of this tool should not be a monolithic application, but a context-aware system. Upon launch, the expert user would select their operative posture: "Institutional" or "Commercial." This choice would fundamentally alter the system's behavior, from its recommendations to its explanations, ensuring the AI's goals are always aligned with the user's business context.

5.4.2 Dual-Purpose Strategic Intelligence

The concept of "Ecosystem Intelligence" becomes far more powerful in this dual-role system. In the Institutional role, the system would analyze the market to identify technology gaps, produce public reports on innovation trends, and advise policy-makers—acting as a true market observatory. In the Commercial role, it would pivot to become a strategic business development tool, identifying the most lucrative or synergistic partnership opportunities for the organization itself.

5.4.3 Adaptive Explainability and Generative AI

The XAI and generative AI features would adapt to the selected role. As a neutral broker, the system would produce highly auditable, transparent explanations focused on demonstrating fairness and neutrality to all parties, and could generate objective summaries for both sides of a potential match. As a commercial partner, the AI would instead highlight the strategic value and business case for the organization, and could produce tailored internal reports or “first-contact” emails tailored to its own business goals.

5.4.4 From Explanations to Patterns

Beyond justifying individual recommendations, explanations can be aggregated over time to surface recurring patterns—what consistently drives successful matches in a given context. In this way, the explanation layer becomes a source of business intelligence for managers, supporting portfolio views and strategic choices rather than only case-by-case validation.

5.4.5 A Simple Post-Match Feedback Loop

A further direction is a light feedback loop to record the real outcomes of partnerships and review them against prior recommendations. Over time, this creates an evidence layer that helps the organization understand which types of matches translate into durable agreements, strengthening decision support while preserving the human-in-the-loop role.

5.4.6 Design Variants Where Appropriate

Future iterations may explore design variants of the matching engine where appropriate, with the goal of improving robustness and coverage without compromising transparency or the simplicity of adoption.

5.4.7 The End Goal: A Dynamic Strategic Partner

This vision culminates in a system that evolves beyond a simple decision-support tool into a dynamic strategic partner. By understanding the user’s role, adapting its analysis, and tailoring its communication, the AI becomes an extension of the expert’s own strategic function. Building such a context-aware system, potentially on a collaborative and privacy-preserving federated architecture as a long-term goal, represents the true future of enterprise AI: a tool that does not just answer questions, but understands the context in which they are asked.

5.4.8 Closing Reflection

This thesis looked at how to balance performance with transparency, innovation with regulation, and efficiency with accountability. The prototype is not a final solution; it simply shows that these trade-offs can be handled in practice under realistic constraints. It illustrates that a privacy-aware, explainable decision-support tool can provide value while keeping human oversight and clear audit trails.

Some questions remain about how organizations will balance imperfect explanations with day-to-day needs and how trust will be sustained at scale. Progress will benefit from iterative technical work alongside governance choices, defined roles, and accountability. The contribution here is a starting point and a direction: AI as a set of tools embedded in organizational processes, supported by ongoing monitoring, evaluation, and refinement.

Bibliography

Adadi, Amina and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.

Arrieta, Alejandro Barredo et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

Bach, Sebastian et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

Bana e Costa, Carlos A. and Jean-Claude Vansnick. "MACBETH: An interactive path towards the construction of cardinal value functions". In: *International Transactions in Operational Research* 1.4 (1994), pp. 489–500.

Belton, Valerie and Theodor J. Stewart. *Multiple Criteria Decision Analysis: An Integrated Approach.* Boston: Springer Science & Business Media, 2002.

Benramdane, Mustapha Kamal et al. "Supervised Machine Learning for Matchmaking in Digital Business Ecosystems and Platforms". In: *Information Systems Frontiers* 26.4 (2024), pp. 1331–1343. DOI: 10.1007/s10796-022-10357-3.

Brans, J. P. and P. Vincke. "A preference ranking organisation method: The PROMETHEE method for multiple criteria decision-making". In: *Management Science* 31.6 (1985), pp. 647–656. DOI: 10.1287/mnsc.31.6.647.

Brella. *The most advanced event matchmaking software.* Product overview: AI-powered, intent-based matchmaking, 1:1 meeting proposals; accessed 2025-09-18. 2025. URL: https://www.brella.io/event-matchmaking.

Brella Help Center. *Introduction — Brella Matchmaking.* How it works: attendees select interests/intents; AI recommendations; accessed 2025-09-18. 2025. URL: https://help-organizers.brella.io/en/articles/177659-introduction.

Buchanan, Bruce G and Edward H Shortliffe. *Rule based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence).* Addison-Wesley Longman Publishing Co., Inc., 1984.

Confalonieri, Roberto et al. “A historical perspective of explainable artificial intelligence”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11.1 (2021), e1391.

Covington, Paul, Jay Adams, and Emre Sargin. “Deep Neural Networks for YouTube Recommendations”. In: *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys ’16)*. Boston, MA, USA: ACM, 2016. DOI: 10.1145/2959100.2959190.

Deck, Luca et al. “A Critical Survey on Fairness Benefits of Explainable AI”. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*. Accessed: 2025-04-06. ACM, 2024. URL: https://facctconference.org/static/papers24/facct24-105.pdf.

DeGrave, Alex J, Joseph D Janizek, and Su-In Lee. “AI for radiographic COVID-19 detection selects shortcuts over signal”. In: *Nature Machine Intelligence* 3.7 (2021), pp. 610–619.

Deloitte. *The Case for Artificial Intelligence in Combating Money Laundering and Terrorist Financing.* Tech. rep. Accessed: 2025-03-04. Deloitte, June 2020. URL: https://www2.deloitte.com/content/dam/Deloitte/jp/Documents/financial-services/bk/en-the-case-for-artificial-intelligence-in-combating-money-laundering-and-terrorist-financing.pdf.

Doshi-Velez, Finale and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

Dyer, Jeffrey H. and Harbir Singh. “The Relational View: Cooperative Strategy and Sources of Interorganizational Competitive Advantage”. In: *Academy of Management Review* 23 (1998), pp. 660–679. URL: https://api.semanticscholar.org/CorpusID:167965404.

Enterprise Europe Network. *About the Enterprise Europe Network.* States largest online database of business opportunities and partnering services; accessed 2025-09-18. 2025. URL: https://een.ec.europa.eu/about-enterprise-europe-network.

— *Partnering opportunities.* Live partnering listings, searchable profiles; accessed 2025-09-18. 2025. URL: https://een.ec.europa.eu/partnering-opportunities.

Fawcett, Tom. “An Introduction to ROC Analysis”. In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

Gaspar, Diogo, Paulo Silva, and Catarina Silva. “Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron”. In: *IEEE Access* (2024).

Google Developers. *Two-Tower Retrieval for Recommendation.* Accessed 2025-09-14. 2022. URL: `https://developers.google.com/machine-learning/recommendation/two-tower`.

Goyal, Navita et al. “The impact of explanations on fairness in human-ai decision-making: Protected vs proxy features”. In: *Proceedings of the 29th International Conference on Intelligent User Interfaces.* 2024, pp. 155–180.

Grant, Robert M. “Toward a Knowledge-Based Theory of the Firm”. In: *Strategic Management Journal* 17.S2 (1996), pp. 109–122.

Grip Events. *How to improve your event networking with AI matchmaking.* Explains app onboarding and data use to refine matches; accessed 2025-09-18. 2024. URL: `https://www.grip.events/news/how-to-improve-your-event-networking-with-ai-matchmaking`.

— *Create more valuable B2B events with smarter matchmaking.* Claims 70M+ yearly recommendations, 16 ML algorithms, and use of platform interactions; accessed 2025-09-18. 2025. URL: `https://www.grip.events/products/event-matchmaking`.

— *Grip - The AI-powered Event Platform Built for Business.* Features include AI matchmaking, meeting management, mobile app; accessed 2025-09-18. 2025. URL: `https://www.grip.events/`.

Gunning, David and David Aha. “DARPA’s explainable artificial intelligence (XAI) program”. In: *AI magazine* 40.2 (2019), pp. 44–58.

Iansiti, Marco and Roy Levien. *The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability.* Harvard Business School Press, 2004.

Ioffe, Sergey and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML).* 2015.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou. “Billion-Scale Similarity Search with GPUs”. In: *IEEE Transactions on Big Data.* Vol. 7. 3. 2019, pp. 535–547. DOI: `10.1109/TBDATA.2019.2921572`.

Kaplan, Robert S. and David P. Norton. "The balanced scorecard: Measures that drive performance". In: *Harvard Business Review* 70.1 (1992), pp. 71–79.

Kazantsev, Nikolai et al. "Investigating barriers to demand-driven SME collaboration in low-volume high-variability manufacturing". In: *Supply Chain Management: An International Journal* 27.2 (2022), pp. 265–282.

Keeney, Ralph L. and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs.* Cambridge University Press, 1993.

Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

Lundberg, Scott M and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

Mazzine, Raphael et al. "Counterfactual explanations for employment services". In: *International workshop on Fair, Effective And Sustainable Talent management using data science.* 2021, pp. 1–7.

Nvidia and Centre for Data Ethics and Innovation (CDEI). *Explainable AI for Credit Risk Management: Applying Accelerated Computing to Enable Explainability at Scale for AI-powered Credit Risk Management Using Shapley Values and SHAP.* Accessed: 2025-03-04. Gov.uk. June 2023. URL: https://www.gov.uk/ai-assurance-techniques/nvidia-explainable-ai-for-credit-risk-management-applying-accelerated-computing-to-enable-explainability-at-scale-for-ai-powered-credit-risk-management-using-shapley-values-and-shap.

OutSail. *How Workday's Acquisition of HiredScore is Transforming HR Tech.* Accessed: 2025-04-04. Nov. 2024. URL: https://www.outsail.co/post/workdays-acquisition-of-hiredscore-reshaping-hr-technology.

Pfeffer, Jeffrey and Gerald R. Salancik. *The external control of organizations: A resource dependence perspective.* Harper & Row, 1978.

Porter, Michael E. *Competitive Advantage: Creating and Sustaining Superior Performance.* New York, NY: Free Press, 1985.

Powell, Walter W. "Neither Market Nor Hierarchy: Network Forms of Organization". In: *Research in Organizational Behavior* 12 (1990), pp. 295–336.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). Official Journal of the European Union. Articles 13–15 and 22. 2016.

Reimers, Nils and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *arXiv preprint arXiv:1908.10084* (2019).

Reuver, Mark de, Carsten Sørensen, and Rahul C. Basole. “The digital platform: a research agenda”. In: *Journal of Information Technology* 33 (2018), pp. 124–135.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “" 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.* 2016, pp. 1135–1144.

Roy, B. “The outranking approach and the foundations of ELECTRE methods”. In: *Theory and Decision* 31.1 (1991), pp. 49–73.

Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

Saaty, Thomas L. “Decision making with the analytic hierarchy process”. In: *International Journal of Services Sciences* 1.1 (2008), pp. 83–98. DOI: 10.1504/IJSSCI.2008.017590.

Sarikaya, Ferhat. *The Quest for Explainability in Artificial Intelligence: Challenges, Progress, and Future Directions.* Accessed: 2025-04-04. 2024. URL: https://medium.com/@ferhatsarikaya/the-quest-for-explainability-in-artificial-intelligence-challenges-progress-and-future-3e36626d58ae.

Schneider, Johannes. “Explainable generative ai (genxai): A survey, conceptualization, and research agenda”. In: *Artificial Intelligence Review* 57.11 (2024), p. 289.

Selvaraju, Ramprasaath R et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision.* 2017, pp. 618–626.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. “Learning important features through propagating activation differences”. In: *International conference on machine learning.* PMlR. 2017, pp. 3145–3153.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

SmythOS. *Top Use Cases of Explainable AI: Real-World Applications for Transparency and Trust.* Accessed: 2025-04-04. 2025. URL: `https://smythos.com/ai-agents/agent-architectures/explainable-ai-use-cases/`.

Spence, Michael. "Job Market Signaling". In: *The Quarterly Journal of Economics* 87.3 (1973), pp. 355–374.

Subasi, Omer et al. "A critical assessment of interpretable and explainable machine learning for intrusion detection". In: *arXiv preprint arXiv:2407.04009* (2024).

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

Teece, David J., Gary Pisano, and Amy Shuen. "Dynamic capabilities and strategic management". In: *Strategic Management Journal* 18.7 (1997), pp. 509–533. DOI: `10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z`.

Transmit Security. *Solving AI's Black-Box Problem with Explainable AI and SHAP Values.* `https://transmitsecurity.com/blog/solving-ais-black-box-problem-with-explainable-ai-and-shap-values`. Accessed: 2025-03-04. Apr. 2023.

Vodafone Group PLC and Nokia. *Automating Anomaly Detection with Machine Learning in Telecom Networks.* `https://www.acceldata.io/blog/automate-data-anomaly-detection-with-machine-learning-in-telecom-networks`. 2023.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.

Wiegreffe, Sarah and Yuval Pinter. "Attention is not not explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics, 2019, pp. 11–20.

Workday. *Responsible AI and Bias Mitigation.* Accessed: 2025-04-04. 2025. URL: `https://www.workday.com/en-us/legal/responsible-ai-and-bias-mitigation.html`.

Yim, Jason et al. *Using AI to predict retinal disease progression*. Accessed: 2025-04-04. Google DeepMind. 2020. URL: `https://deepmind.google/discover/blog/using-ai-to-predict-retinal-disease-progression/`.

Zhang, Jiajin et al. “Revisiting the trustworthiness of saliency methods in radiology AI”. In: *Radiology: Artificial Intelligence* 6.1 (2023), e220221.