



December 2021

HOW DATA SCIENCE
CAN ENABLE

SFDR
REPORTING

QUICK GLOSSARY

CSRD: Corporate Sustainability Reporting Directive

DNSH: Do No Significant Harm

OECD: Organisation for Economic Co-Operation and Development

PAI: Principal Adverse Impact

SFDR: Sustainable Finance Disclosure Regulation

TCFD: Task Force on Climate-related Financial Disclosures

UNGC: United Nations Global Compact

The European Union’s Sustainable Finance Disclosure Regulation (SFDR) could revolutionize sustainability reporting—and, in turn, rescope the data that companies track to measure their ESG performance.

The purpose of SFDR is to improve the transparency of ESG disclosures by financial product and service providers. As not every piece of relevant data is available at scale yet, compliance with SFDR also requires financial market participants to somehow run before they walk. In the intermediary phase, when data providers are in the process of ramping up their offer, this could be counterproductive and result in poor or misreporting. Data science can help.

Making the right data available at scale requires overcoming two main obstacles:

- **Low reliability of reported data.** This can occur because reported company data are fragmented and non-standardized, and conflicting or unreliable values exist across different providers.
- **Incomplete data coverage of metrics and industry sectors.** This occurs because of partial or nonexistent reporting.

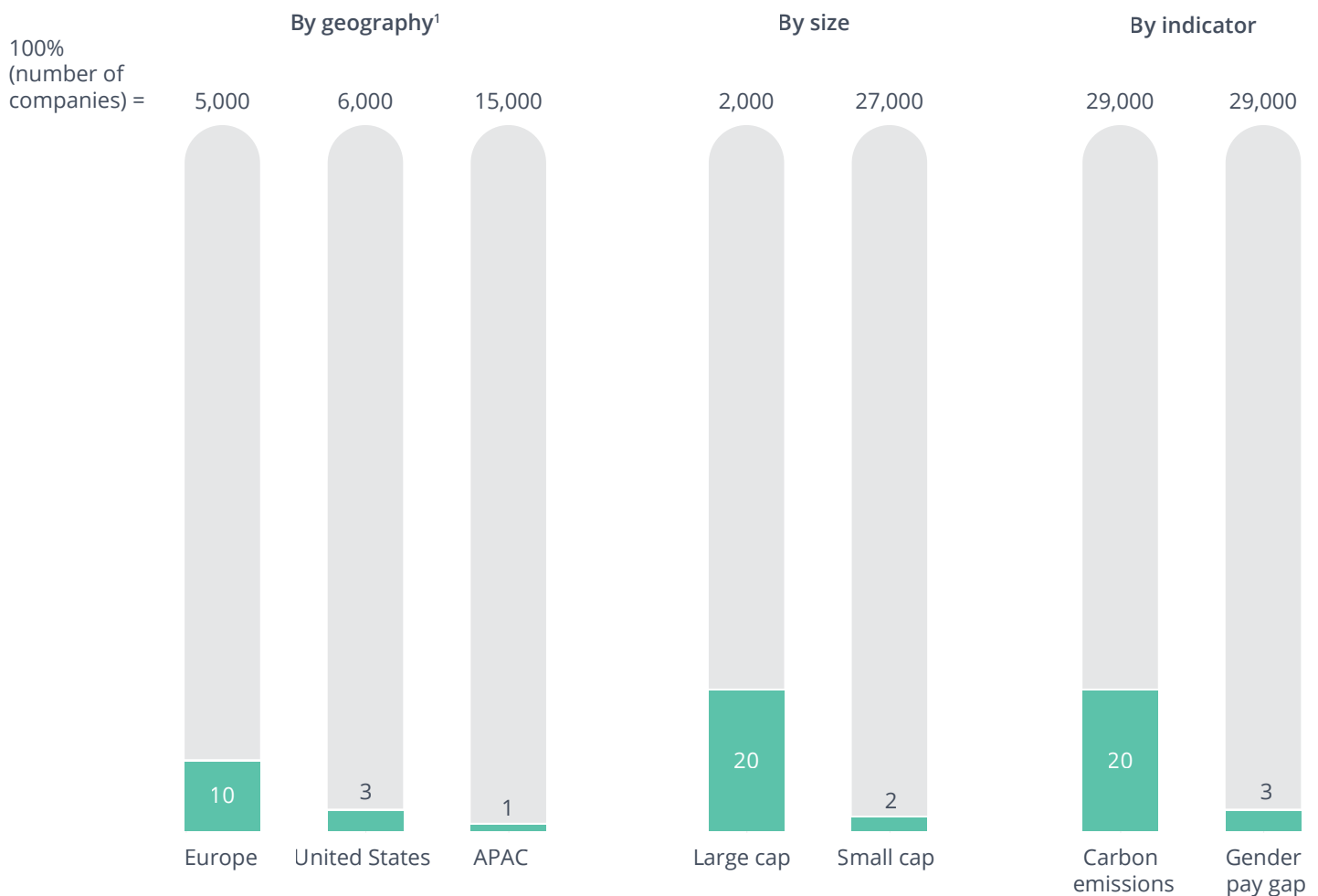
Figure 1 illustrates SFDR principal adverse impacts (PAI) coverage gaps by geography, size, and selected indicators based on a Clarity AI data science-enabled analysis of 29,000 companies.¹ Globally, only 3% of companies analyzed reported more than 70% of the 14 mandatory PAI. Europe leads the way with 10% of firms meeting this

coverage threshold, while just 3% of US firms and 1% of APAC firms reported the same. One in five large-cap firms met the threshold, but just one in 50 small-cap firms did. Coverage also varies widely by indicator: 20% of the 29,000 firms disclose carbon emissions data, but just 3% do the same on gender pay gap data.

FIGURE 1

Our analysis indicates significant coverage gaps in SFDR principal adverse impacts (PAIs).

Companies reporting more than 70% of PAIs (n = 29,000 companies), %



1. Not all geographies are represented.

1. For key SFDR PAI, our data science expertise has allowed us to multiply data coverage by approximately five times, on average. In some cases, when no data were available, we developed entire new data sets. Alongside coverage issues, financial market participants also face reported-data reliability issues, which is another area where our data science approach provides significant improvements.

In addition to meeting the PAIs, SFDR requires evidence of inclusion of good governance practices and that companies account for the Do Not Significant Harm (DNSH) test—each of which comes with its own data coverage challenges. The implementation of the European Commission’s Corporate Sustainability Reporting Directive (CSRD) will help bridge the gap, eventually compelling close to 50,000 companies to report sustainability performance on a comprehensive set of metrics. CSRD will be fully implemented by 2025, and non-European jurisdictions are likely to lag even further behind.

CLARITY AI IS A SUSTAINABILITY TECH FIRM AND PLATFORM WITH THE MISSION OF BRINGING SOCIETAL IMPACT TO MARKETS.

SFDR reporting requirements add another layer of data to the 200 metrics that Clarity AI already provides to evidence performance on ESG risk and impact on the world, as well as alignment with climate targets (including those of the Task Force on Climate-related Financial Disclosures) and alignment the UN’s Sustainable Development Goals (SDGs). As a one-stop shop, Clarity AI also provides clients with robust and comprehensive solutions to meet their SFDR disclosure and product-design requirements, leveraging our data science capabilities.

In this paper, we provide specific examples of the merits of our data science approach for SFDR reporting. We address the prerequisite for sound modeling and recommendations on how to use the data. We also highlight current limitations and how we intend to further develop the SFDR analysis and reporting module in the coming months. The paper is framed around three specific use cases leveraging different data science techniques,

illustrated through SFDR requirements:

1. How data science can improve reliability of reported data
2. How machine learning can expand data coverage
3. How natural language processing can inform metric development


Whereas SFDR covers several asset classes, including sovereign bonds and real estate, we will focus on corporates.

FIGURE 2

Clarity AI achieves optimized outcomes thanks to three key differentiators.

- 01 DATA SOURCES**

 Assemble the largest collection of structured and unstructured sources to cover all key topics and industry sectors
- 02 TECHNICAL DATA EXPERTISE**

 Aggregate, clean, and standardize assembled database to improve data quality, and continuously improve models through in-house and external expertise (partnerships with academia and consulting firms)
- 03 ARTIFICIAL INTELLIGENCE**

 Implement state-of-the-art machine learning and data science techniques with scalability in mind for automatic best source selection and to obtain accurate estimates for non-reported data, increasing reliability and coverage

TO WHAT EXTENT CAN MODELED DATA BE USED IN SFDR REPORTING?

The first question on our journey was to figure out to what extent SFDR allows financial market participants to use modeled data. To answer this fundamental question, we reached out to our sustainable regulation external advisor and partner Eco:Fact.

One of the key aims of the SFDR, and other sustainable finance regulations, is to reduce an asymmetry of information between financial market participants and investors. Consequently, financial market participants are expected to support their reporting and make decisions based on data that still might not be unavailable. For example, this can be noted when the EU's innovative sustainable finance regulations introduced requirements for reporting on sustainability risks and adverse impacts on sustainability factors.

Although these two categories are closely related, they require financial market participants to assess sustainability topics, such as climate change and human rights violations, from different perspectives:

- a “sustainability risk” analysis focuses on potential material negative impacts on the value of an investment that stem from sustainability factors (e.g., the impact of sea level rise on property values).
- consideration of “adverse sustainability impacts” concentrates on an investment's negative effects on sustainability impacts (e.g., investment in highly polluting companies that negatively impact ecosystems and individuals' health).

Appropriate data is needed to conduct assessments such as those described above; data availability, accessibility, and reliability are

central to financial institutions' efforts to answer questions about sustainability risks and adverse impacts, and thus meeting the SFDR's expectations. Regulators are aware of the data-related challenges institutions face, and they are currently designing solutions to bridge this data gap via, for example, the proposed Corporate Sustainability Reporting Directive (CSRD). In the meantime, the European Union regulator provides tools that financial market participants can use to address the issue of data availability.

One tool is mentioned in Article 7(2) of the Regulatory Technical Standards (RTS) (commonly referred to as level 2). (The RTS under the SFDR are expected to become applicable sometime in 2022.) This provision is relevant for situations where financial market participants are requested to disclose data on principal adverse impact indicators but that data is not readily available. In this context, financial market participants are expected to use “best efforts” to obtain the information they need, either directly from investee companies or by carrying out additional research. They can also cooperate with third-party data providers or external experts or make reasonable assumptions. It should be noted that financial market participants must also report what constitutes their best efforts.

In this situation, financial market participants' use of modeled data is one solution to tackle the challenge of data gaps—this strategy fulfills the criteria of the expectation to use best efforts. For example, modeled data enables financial market participants to base their disclosures on an approach that is verifiable and that is used by other market participants to make reasonable assumptions about impacts on sustainability factors.



How data science can improve reliability of reported data

Sustainability performance data are still in their early days. The CSRD will eventually make these data part of companies' annual reports with third-party auditing. However, CSRD will not be fully implemented until 2025, and in the meantime, limited reliability in reported data is to be expected.

This limited reliability applies even to broadly used quantitative metrics such as Scope 1 CO₂

emissions, which—despite being a highly material metric—can suffer from high variability among data sources due to errors, lack of standardization, and overall poor data quality from provider to provider. This is true even when dealing with data reported by the companies themselves. The higher the variability, the less reliable the data.

Consider the example variability analysis in Figure 3.

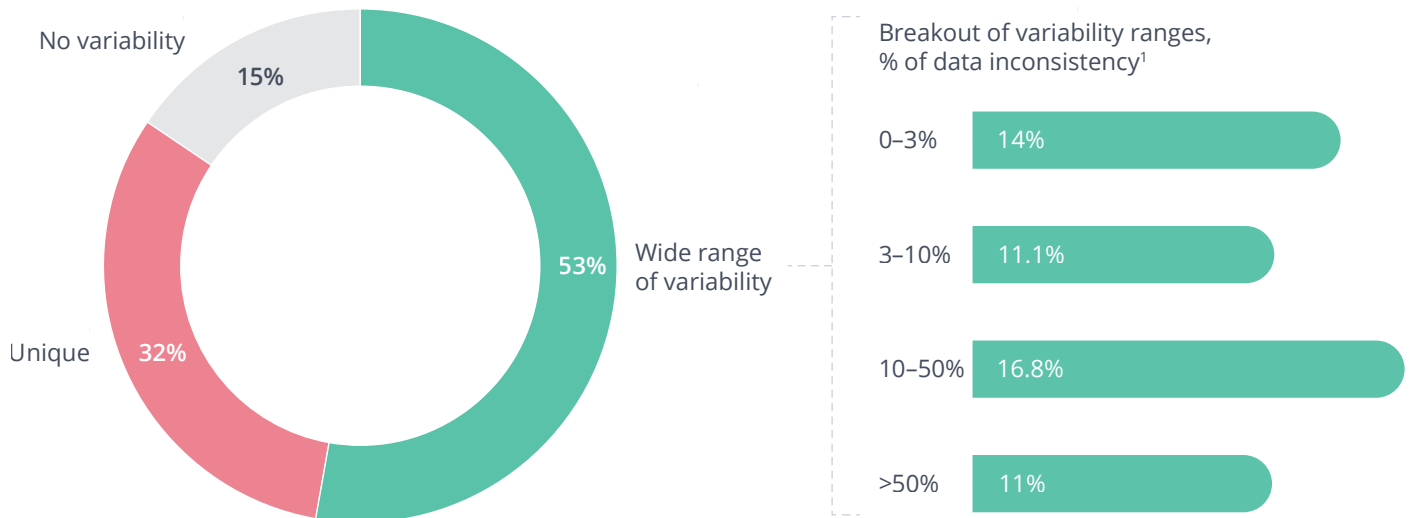
Key finding: The variability among data providers used by Clarity AI for Scope1 CO₂ emissions, a high materiality metric, is significant.

- One in four providers sees data inconsistencies of 10% or higher in the reported values, according to different providers.
- For 11% of providers, variability exceeds 50%.
- Only 15% of the data show no variability—that is, full convergence.

FIGURE 3

Data variability among data providers for Scope 1 CO₂ Emissions reported by companies.

Variability among different data providers used by Clarity AI for Scope 1 CO₂ emissions



1. Figures may not sum to 100%, because of rounding.

CLARITY AI'S APPROACH TO IMPROVE RELIABILITY

Clarity AI leverages three key differentiators to establish the most reliable database available today (Figure 4).

- We assemble the largest collection of structured and unstructured data sources in a global database.
- We use in-house and external technical data expertise to aggregate, clean, and standardize this database.
- We leverage proprietary machine-learning algorithms and data science techniques to detect outliers and automatically select the best source for overlapping data, as well as to obtain accurate estimates for non-reported data.

Key differentiator: Data sources

Clarity AI draws on **more than 50 generalist and specialist external data providers**, for a total

of more than two million data points of various types (for example, quantitative, qualitative, and news). Clarity AI also has **proprietary data** from machine-learning models that estimate metrics to complement organizations' non-disclosed information.

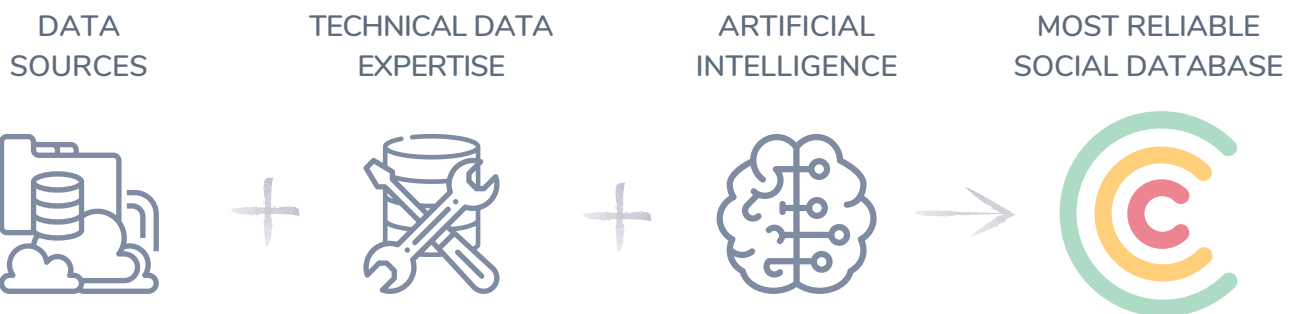
Finally, **exclusive data sources** from Clarity AI partnerships with worldwide, recognized data providers (for example, for controversial news) allow deeper and richer insight generation.

Key differentiator: Technical data expertise

Clarity AI's data engineering and DevOps teams are experts in data life-cycle management, and they leverage bleeding-edge technology and tools for automated data ingestion, processing, validation, and storage. Our team can expertly clean and standardize a company's other data, classifying it into peer groups and identifying key operating metrics.

FIGURE 4

Clarity AI has the most reliable social database today for SFDR reporting.



Key differentiator: Artificial intelligence

Confirmed data are great; triple-confirmed data are better. Clarity AI uses its multiple sources, as well as overlapping coverage of key metrics, to ensure data consistency and reliability.

To remove potential inconsistencies within this consolidated database, Clarity AI's proprietary machine-learning algorithms choose the best sources and detect outliers just as an analyst would do based on domain expertise—but at scale and without human bias.

For example, for key quantitative data such as CO₂ emissions, water, and waste, overlapping coverage from multiple providers is over 70%. In these cases, machine-learning algorithms choose the best sources from those available. For the rest of the quantitative metrics, analytics algorithms are used for outlier identification.

Overlapping coverage is also crucial for metrics such as policy and controversy data, where an absence of evidence can be misunderstood as evidence of absence. For such metrics, Clarity AI mines sources for overlapping coverage of more than 80%.

As inputs, the algorithms use historical data trends, comparison against industry peers, variability among data providers, and expected value range based on proprietary estimation models.

CASE STUDY

SALESFORCE

The number for Salesforce's 2019 Scope 1 CO₂ emissions was reported inconsistently in a variety of data sources. Two data providers offered a value of 5,800 tons. A third provider said 5,000 tons, and a fourth reported 50,000 tons. Clarity AI's algorithm concluded that the 5,000-ton value was the most reliable, and this conclusion was then backed up by Salesforce's own annual report.

The Salesforce logo is displayed in a light gray, lowercase font within a white, stylized cloud shape. The cloud has three main lobes and is set against a light beige, textured background that resembles a piece of torn paper.

CLARITY AI DATA ARE 100% RELIABLE

Applied at scale, Clarity AI's algorithms deliver a significant increase in data reliability. Consider the example of Scope 1 CO₂ data: Figure 5 shows a comparison of the coverage and data quality of five different data providers against the dataset provided by Clarity AI. For each provider, we analyzed whether data are even available—and, if they were, whether the data are reliable.

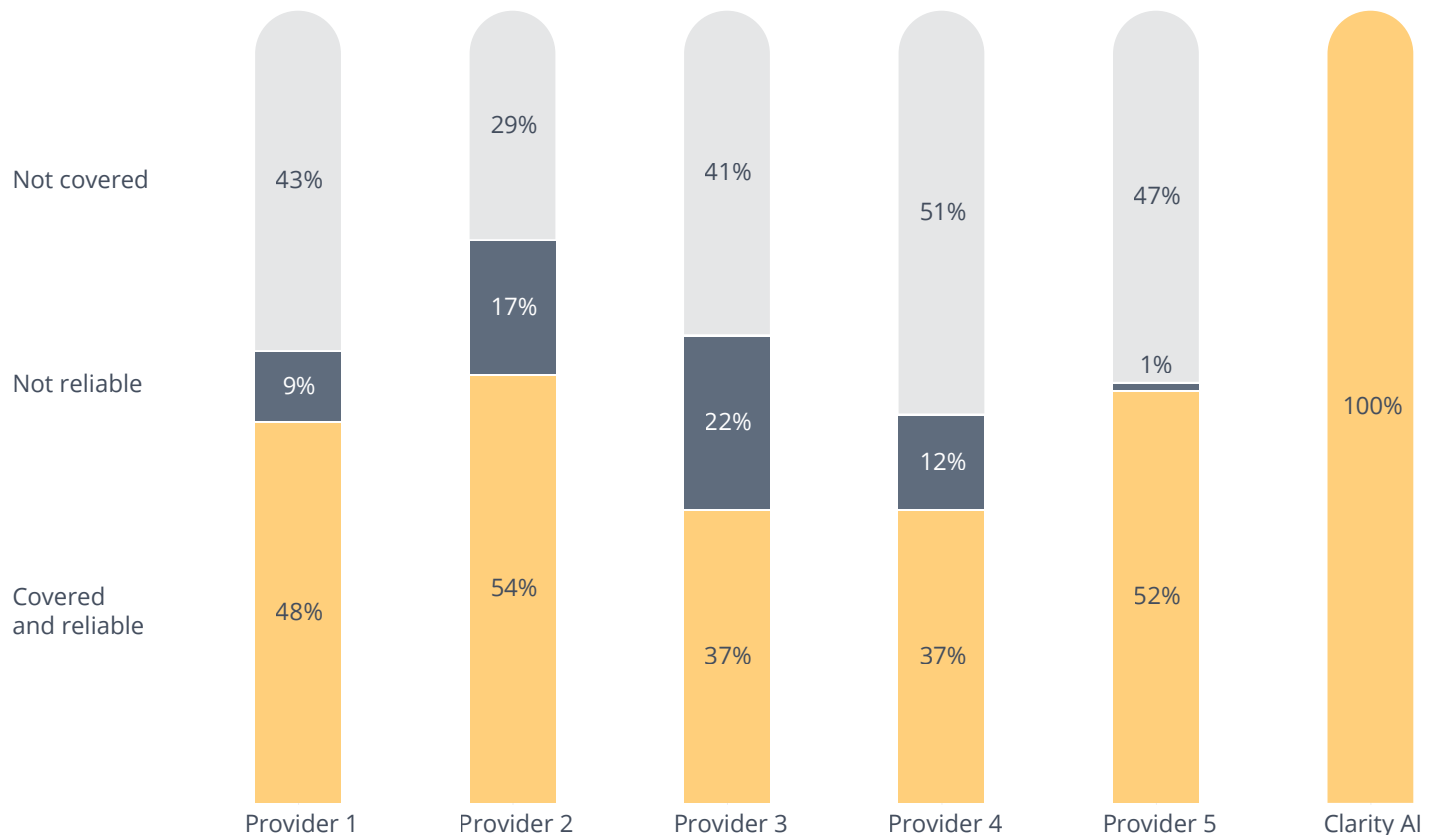
Key finding: The Clarity AI dataset is both larger and more reliable.

- No provider comes close to the 100% reliable data coverage provided by Clarity AI.
- Only 54% of the second-place provider's data are both covered and reliable.
- Up to 51% of other providers' data are not covered.
- Up to 22% of other providers' data are unreliable.

FIGURE 5

Compared with other data providers, Clarity AI offers a significant increase in data reliability and coverage.

Comparison of the Clarity AI database for Scope 1 CO₂ emissions with specialized data providers, % of datasets



02



How machine learning can expand data coverage

Lack of data coverage is a major hurdle that can be overcome through the use of machine learning. Today, 80% of listed companies do not report required sustainability data. That means that, regardless of reliability issues, only 20% of publicly listed companies report comprehensive data on sustainability as a baseline.

As demonstrated above, many providers may then pile on partial or missing information, making it

difficult to create consistent scores across peers and potentially skewing scores toward companies that disclose selectively by leaving out data on indicators for which they are behind. For this reason, Clarity AI leverages available company information and machine-learning algorithms to fill in the information gaps to give the fullest available picture.

Geographically, Europe has been leading the way with national regulations on climate change reporting for corporations, which crystallizes in the highest GHG reporting coverage among major world regions (Figure 6). Meanwhile, the US Securities and Exchange Commission is preparing a specific climate reporting regulation for 2022. The expectation is that reporting in North America will catch up to the rate in Europe within the next couple of years.

POOR PERFORMERS BY INDUSTRY

While the average reporting rate stands at a low 17%, sectors like mining, utilities, oil and gas, airlines, and automobile manufacturers lead the way with reporting rates ranging between 50% and 75%. With reporting rates below 17%, fertilizer and agricultural-product producers are lagging behind—which is especially worrisome given the crucial role the agricultural sector plays in the fight against climate change.

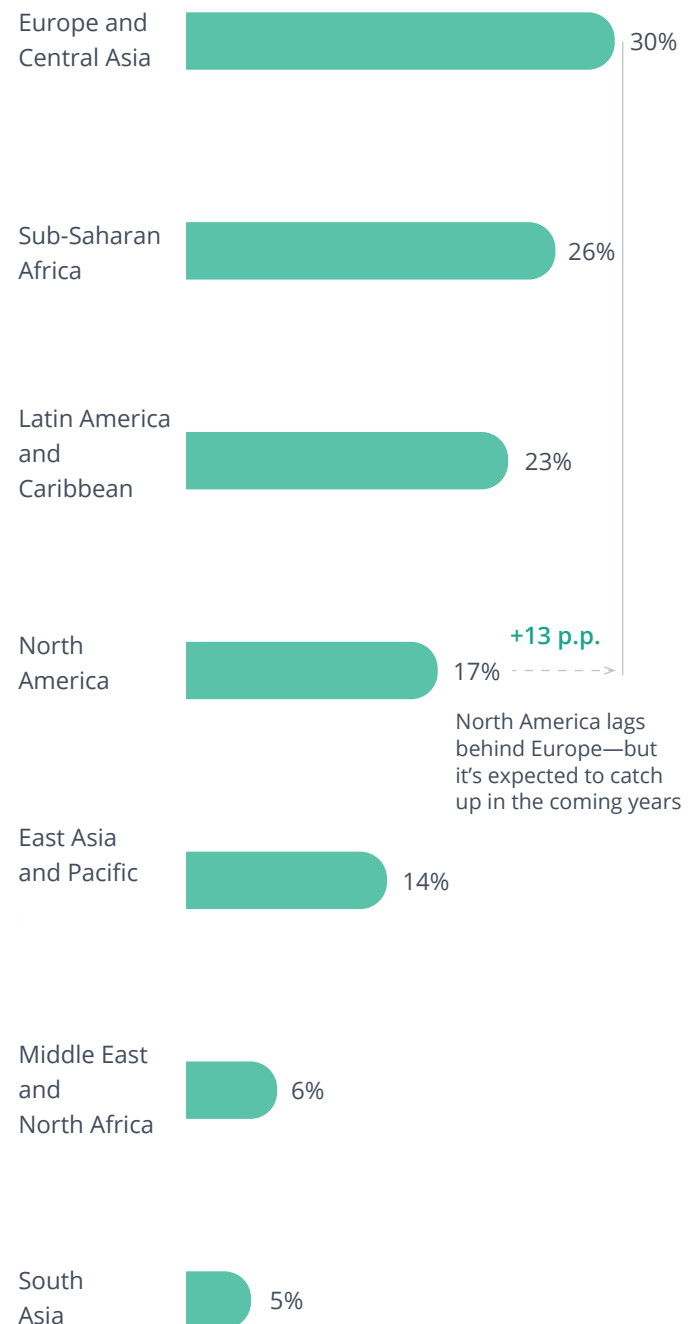
WHAT INDICATORS ARE REPORTED TODAY?

Since 2015, the Task Force on Climate-related Financial Disclosures (TCFD) and various national regulations have supported the development of climate footprint data by companies. It is therefore not surprising that climate footprint data are the most reported among the PAI related to the environment. Conversely, energy use and hazardous waste can be considered very sector specific, partly explaining the low level of reporting.

FIGURE 6

Europe has the highest GHG reporting coverage among major world regions.

GHG Scope 1 emissions reporting coverage by region, n = 29,000 companies



CLARITY AI'S ESTIMATION MODELS

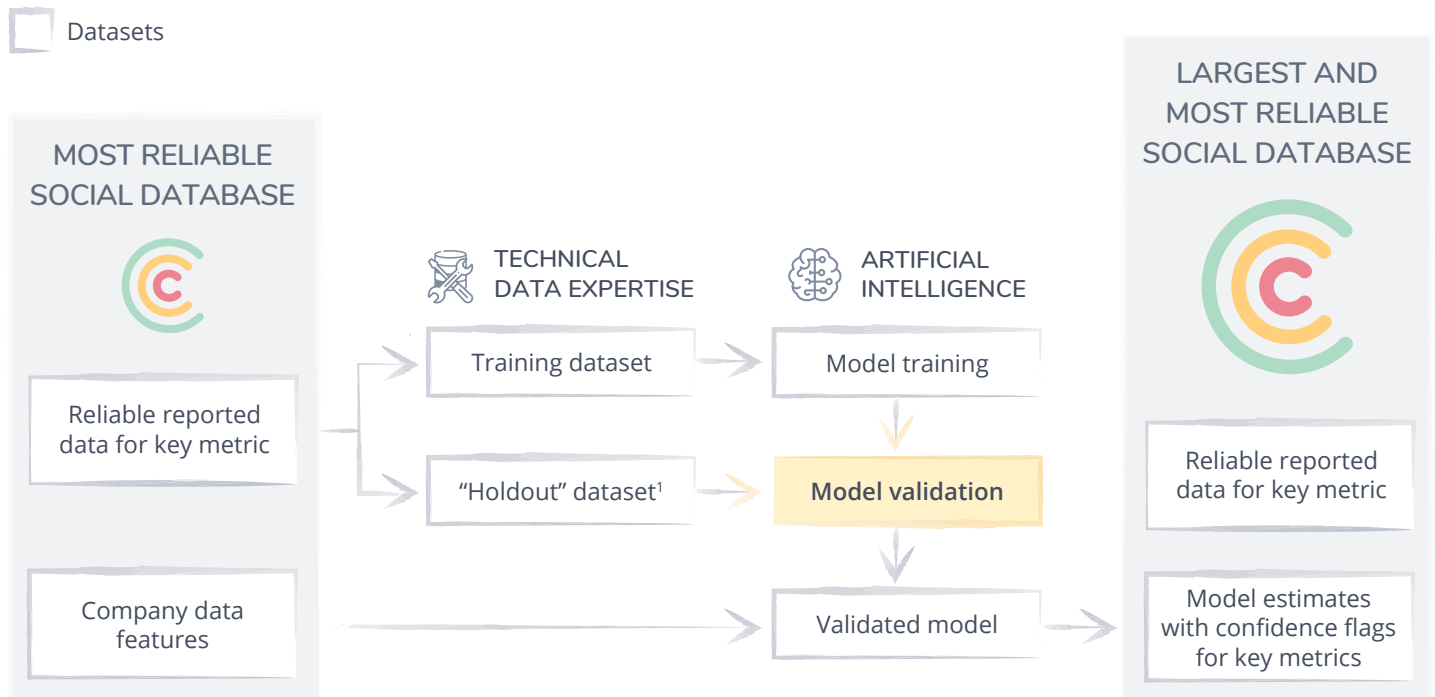
One application of machine learning is our estimation models. The underlying principle of the models is to figure out how sustainability performance metrics can be derived from other corporate attributes. A wide range of both data sources and features (information about the organization, described as “company data features” in Figure 7) are used as input for the estimation models, including, for example:

- What industry are you in?
- What types of products and services do you sell?
- Are you a manufacturer?
- Where do you make your products?
- Where do you sell your products?
- What are your labor costs?
- What are other environmental features that may be correlated with the metric of interest? (This depends on the metric.)

FIGURE 7

Clarity AI employs technical data expertise and artificial intelligence to inform its estimation model process.

Flowchart of estimation model process



1. The holdout subset is used as a clean proxy of the data to be estimated, and allows us to verify the true predictive power of the model.

Key differentiator: Intensity

To begin, Clarity AI's models estimate the intensity of the metric when appropriate (for example, when working with CO₂ emissions, the model will estimate the raw value of the metric divided by the company's revenue in millions of dollars). This strategy has two benefits with respect to simply estimating raw data (for example, CO₂ emissions):

- Intensity (that is, ratios of raw value to revenue in dollars) is better aligned with the concept of efficiency and resource usage.
- Intensity has a more compact range of possible values, improving the performance of a regression model.

A higher intensity can mean a more complex road to metric improvement. Company features can increase or decrease estimate value. For example, a company in diversified metals and mining—a subsector that tends to have higher metric intensity values—can reduce that intensity by shifting to power produced from low-emitting CO₂ technologies. Meanwhile, a construction and engineering company may weigh the pros and cons of pursuing revenue from subsectors with higher intensity metrics, such as industrial services and manufactured products.

Key differentiator: Holdout data

In the holdout methodology, a subset of the available data (usually 20%) is set aside to test the

predictive accuracy of the model. This subset is used as a clean proxy of the data to be estimated and allows Clarity AI to verify the true predictive power of the model, testing it with information from companies that weren't used to train the model.

The main metric used to check how well the model fits is a rank order metric using the Spearman test. This metric has been chosen because it is a good proxy for robust, best-in-class intensity scores. The Spearman test focuses on validating how well the estimates are sorted versus the ground truth values. In other words, it checks that the model predicts higher intensities for companies that present higher intensities and lower intensities for companies that present lower intensities.

As an example, if the model predictions are in the same order as the actual (holdout) data, the ranking metric would have a maximum value of 1. If the predicted order of companies were sorted in random order, the metric would be 0, and in inverse order, the metric would be -1.

On top of ensuring that the model can order companies correctly within each industry, we also look into whether the shape of the distribution of estimated values is similar to the distribution of reported values. This is done by taking into account the order of magnitude of the different quartiles.

Key differentiator: Nonlinearity and interactions

Last, our machine-learning-based estimation models allow us to account for both non-linear and interaction effects, as these are crucial for estimating certain sustainability metrics as CO₂ emissions. On the one hand, as a measure, metric tons of CO₂ per dollar might decrease with the number of employees (due to the efficiencies of economies of scale), but this effect is non-linear and diminishes. On the other hand, emissions per

worker can vary across geography, and thus the interaction between the number of employees and country features is determinant.

CLARITY AI EXPANDS DATA COVERAGE BY 500%

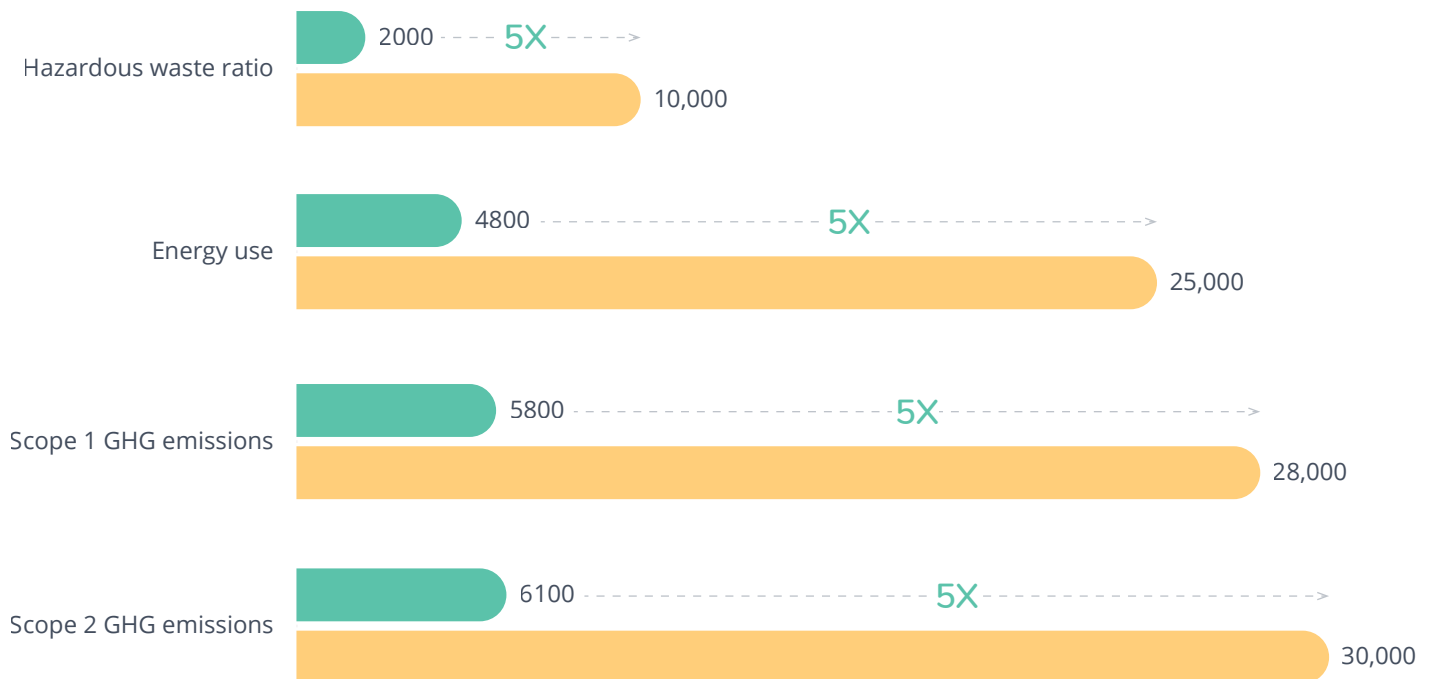
A key feature of Clarity AI is the breadth of our data coverage. Based on our universe of 29,000 companies, our estimation model considerably expands the percentage of companies disclosing environmental PAI (Figure 8).

FIGURE 8

Through its estimation model, Clarity AI offers a fivefold increase on reported data coverage for specific PAI.

Breakdown of data coverage by PAI

● Reported coverage ● Clarity AI's enhanced coverage



Note: Figures may not sum, because of rounding.

03



How natural language processing can inform metric development

Most of the solutions available in the market for the analysis of companies' behavior rely heavily on the manual assessment of news or on sentiment analysis, leading to the following weaknesses: subjective assessment, volume limitation, and limited ability to interpret metrics. These limitations reduce the capability to provide timely and meaningful information in a consistent and transparent way.

CLARITY AI'S CONTROVERSY SCORING SYSTEM

Clarity AI addresses these weaknesses and limitations with a controversy scoring system.² We have built scores using a global news monitoring service as our main source of data, which provides

us with access to a universe of more than 8,500 media publishers that cover 200 countries, with 100,000 new articles added per day from more than 33,000 sources. This adds up to approximately 70 million articles related to the Clarity AI company universe for the last three years.

Our controversy scoring system breaks down into four major steps:

1. incident detection
2. incident classification
3. incident severity scoring
4. event severity scoring

2. We define "controversies" as conflicts, even if just alleged, between a company and any given social agent or stakeholder (for example, employees, inhabitants of a region, or governmental authorities) originated by the breaking of a global norm linked to responsible business conduct, resulting in a risk increase for the issuer.

We can identify the evolution of controversial incidents over a timeline—as well as their severity—for a given company in a specific category from among the 39 categories we assess. For example, a company like Tesla has thousands of articles written on it (23,189 from summer 2017 to winter 2020). Out of those, 3,767 articles are relevant to the business ethics category, but they vary substantially in how severe they are. The AI model considers all relevant articles for each category, then using severity as one key proxy of PAI breach (see details in deep dive that follows).

We consider an “event” to be the whole three-year series of incidents that refers to a specific company in one ESG controversy category. The event score is then calculated through the combination of the resulting maximum severities for the most relevant

incidents within the event. As an output, we obtain an overall score at the company-category level.

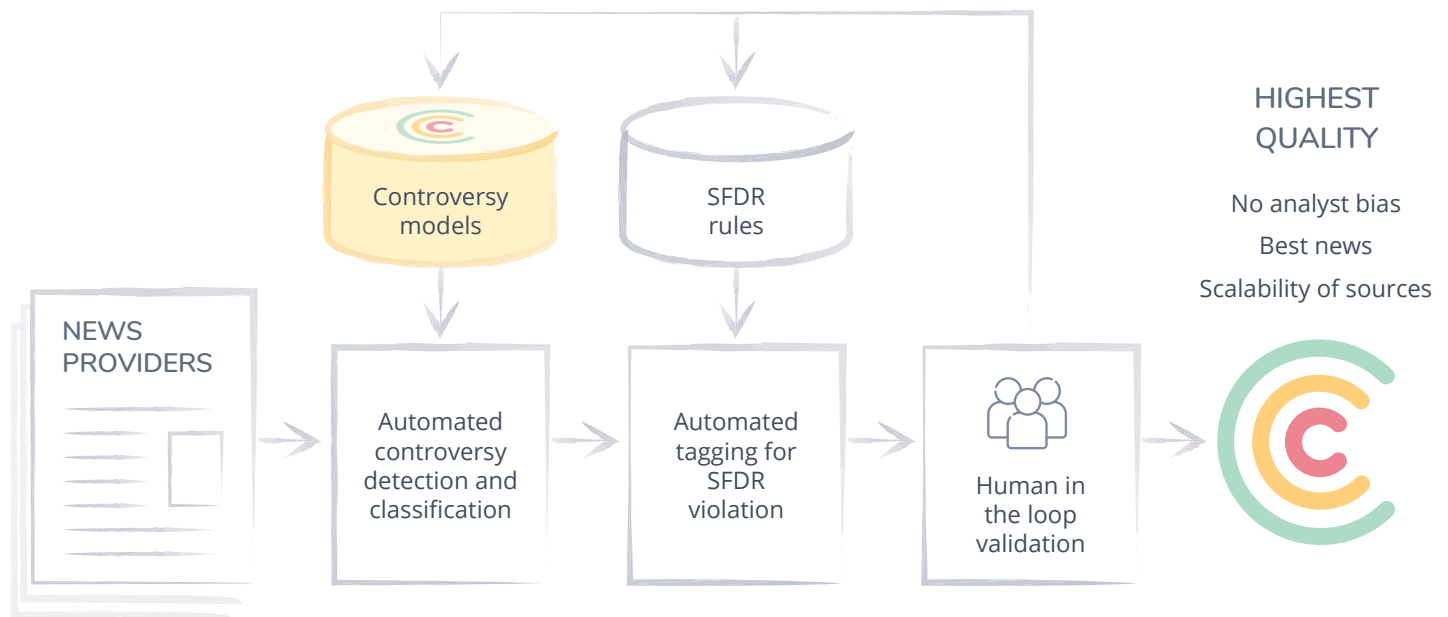
Each of the steps relies on Clarity AI’s proprietary artificial-intelligence models, which are purposefully designed to detect, classify, and assign the corresponding severity. These algorithms are the key factor for our objective analysis, having been trained on subject-matter-expert intelligence through a human-in-the-loop process, with a selection of more than 30,000 articles covering all controversy categories and controversy levels, allowing the model to learn the relevant criteria to be considered in each step.

The combination of our controversy methodology and SFDR rules is illustrated in Figure 9.

FIGURE 9

Clarity AI’s controversy models aid in the assessment of incidents involving SFDR rules.

Breakdown of incident assessment engine



Deep dive: How Clarity's AI's controversy scoring system can identify breaches in biodiversity PAI

While developing solutions for SFDR, we came to the conclusion that for biodiversity and norms-violations PAI, our controversy methodology could provide far more accuracy and reactivity than any existing analyst-based process.

A breach of the Biodiversity PAI indicator implies that a company is negatively affecting—through its activities, services, or products—biodiversity within one of the defined protected biodiversity-sensitive areas. The PAI covers the Natura 2000 network of protected areas, UNESCO World Heritage sites, and the key biodiversity areas (KBAs), as well as other protected areas as specified by the European Parliament and the Council.

To identify breaches of the Biodiversity PAI, we use Clarity AI's Controversies and Exposures Modules to target relevant biodiversity-related issues affecting corporations, as well as to incorporate specialized external sources to ensure wide-ranging coverage.

As a first step, our controversy models, running on natural language processing, help us to identify existing controversies which fall into environmental pillar categories, potentially relevant to the Biodiversity PAI.

As a second step, we further refine the information using specific criteria:

1. Evidence of company involvement in sensitive services, products, and activities (e.g., oil and gas, shale energy, or palm oil).
2. Companies with severe controversies³ in ESG categories: land and biodiversity impact, products environmental impact, waste management impact, product climate change impact, and water use.
3. Occurrence of biodiversity-related keywords, linked to effective negative effects on biodiversity, or to locations within the Natura 2000 network, UNESCO World Heritage sites, and KBAs.

This process helps us identify what could be considered potential breaches of the PAI, and it is then complemented by the intervention of a human in the loop who certifies that:

- a breach has effectively taken place in the recent past (Clarity AI considers up to four years as recent enough)
- the breach is directly located in a protected area or in such close proximity that it would have been affected by the breach—in order not to flag companies unnecessarily

3. During validation, severity is taken into consideration, which encompasses the magnitude of the issue, issue management, business risk, and reputational risk of the company in question.

EXAMPLE 1

Company: Endesa, S.A.

Location: Osona, Northern Catalonia, Spain

Directive: The Birds Directive, Directive 2009/147/EC (SPA)

In the news: “Spain’s Endesa power firm sued over electrocution of birds”

Clarity AI’s approach: The algorithm detected the company of Endesa S.A. as being in a potential breach of the Biodiversity PAI due to:

- involvement in sensitive activities, such as coal power generation, fossil fuel, and nuclear energy production
- recent controversies in land use and biodiversity topics, which indicated negative impacts on biodiversity sensitive areas

To assign a breach label to the company, we consider the magnitude of the issue, the directive, and the location associated and the impact on biodiversity:

- *Magnitude of issue.* The company is being prosecuted over deaths of hundreds of birds—of an endangered species—that have been

electrocuted due to failure of the company to comply with regulations designed to protect wildlife. The company has been accused of environmental crime and crime against wildlife protection. Even though the company presented the Catalan regional government with mitigation plans in 2013, it still has not made the necessary adjustments to its lines, despite repeated warnings.

- *Date.* The lawsuit claims that the incidents happened between 2018 and 2020.
- *Directive and location.* The Birds Directive, Directive 2009/147/EC, covered under this PAI establishes the creation of Special Protection Areas (SPA). Endesa’s actions took place in many of these areas in Spain.
- *Biodiversity impact.* Several endangered as well as migratory species have been affected due to the lack of measures taken by Endesa.

Breach: Yes. This is a clear breach, as the company is being sued due to inaction and continued negative effects, and the birds affected are protected by The Birds Directive.

EXAMPLE 2

Company: Norsk Hydro ASA

Location: Ilha de Marajó, Pará Amazon, Brazil

In the news: “Brazil Group sues Norsk Hydro over alleged pollution”

Clarity AI’s approach: The algorithm detected the company of Norsk Hydro ASA as being in a potential breach of the Biodiversity PAI due to:

- detecting keywords related to a biodiversity-protected area (the Amazon)
- recent controversies in emissions, effluents, and waste topics, which indicated negative impacts on biodiversity-sensitive areas

To assign a breach label to the company, we consider the magnitude of the issue, the directive, and the location associated and the impact on biodiversity:

- *Magnitude of issue.* A group-action lawsuit in the Netherlands has been brought up against the company due to alleged toxic-waste pollution

in northern Brazil, indicating the international reporting on this controversy. This caused health issues for local communities living in the Pará region, as well as environmental issues and pollution within and in close proximity to protected areas.

- *Date.* The spill occurred in 2018.
- *Directive and location.* The damage has affected biodiversity-sensitive areas protected under the KBA Directive.
- *Biodiversity impact.* Due to the nature of the contamination, including water contamination due to leakage of toxic waste containment basins, air pollution, and discharge of soot, several KBAs of the region have been affected.

Breach: Yes. This is a clear breach, as there is an ongoing group-action lawsuit against the company due to its negative impact on the environment and on key biodiversity areas protected by the KBA Directive.

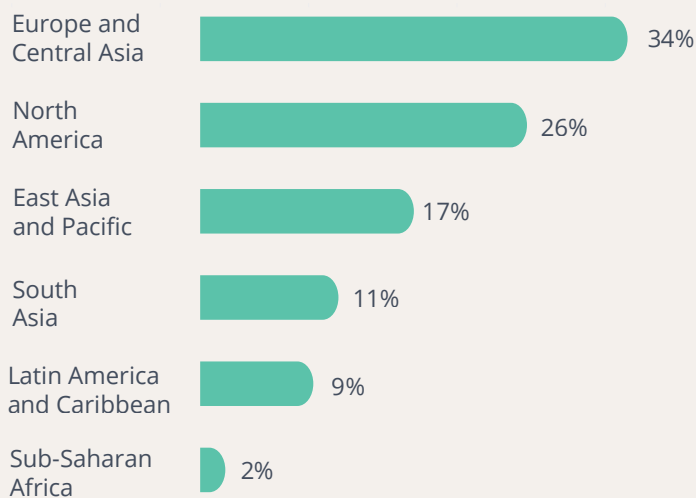
SECTOR AND GEOGRAPHY BREAKDOWN OF CORPORATIONS WITH BIODIVERSITY NEGATIVE IMPACT

Analyzing the outcome of our biodiversity breach analysis, we find that European and North American companies together account for 60% of all adverse biodiversity impacts. Asian companies represents 28%. Paradoxically, Latin American and African companies only account for 11% of total breaches, whereas almost 40% of the biodiversity hotspots are located on these continents. These figures are to some extent driven by the higher share of listed companies in developed versus emerging markets, and they are an indication that there might be some geographical bias in the news media—a topic we will investigate in the future.

FIGURE 10

Despite being the location of 40% of biodiversity hotspots, African and Latin American companies account for only 11% of biodiversity breaches.

Region of origin of companies with adverse biodiversity impact¹



Note: Figures may not sum to 100%, because of rounding.

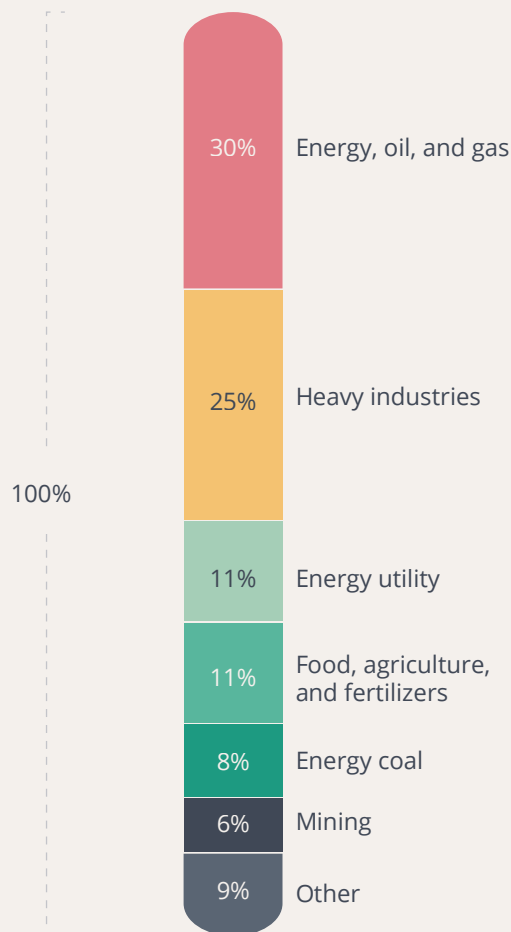
1. 100% = the total number of companies with biodiversity breaches.

When it comes to the sectoral breakdown, energy companies represent the majority of the breaches, with oil and gas companies being the most prominent (30%). Heavy industries represent a quarter of total adverse impacts. The food and agriculture industry is also identified as a significant negative contributor (11%), while mining represents 6% of the total, in line with its share of global GDP (6.9%).

FIGURE 11

Energy companies represent nearly a third of biodiversity breaches.

Sector breakdown of biodiversity breaches¹



1. 100% = the total number of companies with biodiversity breaches.

Deep dive: How Clarity's AI's controversy scoring system can identify norms violations

The different controversial conducts derived from the analysis of global norms have been structured with regard to the criteria of completeness (holding all the topics that can be relevant to investors) and exclusiveness (no significant overlap between categories), resulting in 39 categories (see Annex).

To come up with a specific approach to United Nations Global Compact (UNGC) and Organisation for Economic Co-Operation and Development (OECD) guidelines violations under SFDR PAI #10, we have developed a specific mapping of our controversy categories against the 10 principles

contained in the UNGC as well as against the over 50 policies and recommendations addressed by the OECD guidelines. In this process, the focus was set on the principles and policies that the companies need to adhere to (omitting the very general ones and those that they are only encouraged to follow). The criteria employed for the flag of a PAI violation is for the entity to have a severe or very severe controversy—as defined by the controversy criteria above—in any of the mapped categories to trigger a PAI violation. The recent involvement in the breach of such a norm should also be verified by a human in the loop.

Figure 12 presents an overview of the mapping outcome.

FIGURE 12

Clarity AI's controversy categories can be mapped onto the list of policies addressed by UNGC principles.

Category	UNGC principles
Human rights	1, 2
Labor rights	3, 4, 5, 6
Remuneration and working conditions	
Water use	
Waste management impact	
Products environmental impact	7, 8, 9
Product climate change impact	
Climate change mitigation	
Land & bio impact	
Corruption incidents	10

SECTOR AND GEOGRAPHY BREAKDOWN OF CORPORATIONS VIOLATING UNGC NORMS

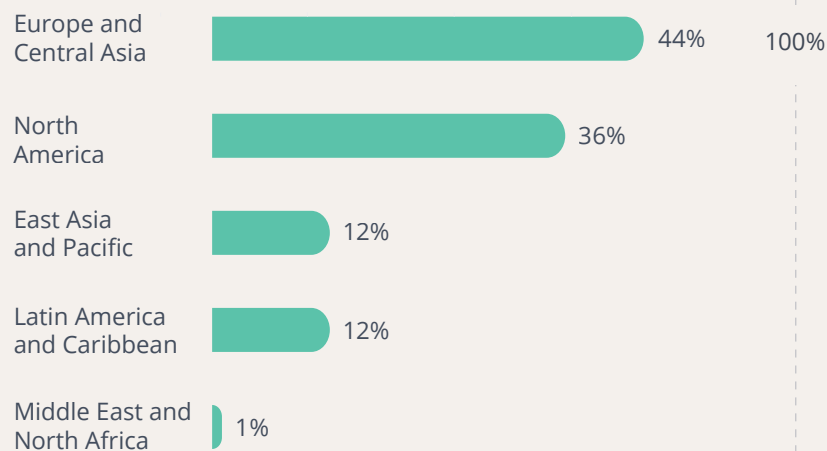
European and American corporations represent 80% of UNGC norms-violation cases identified, thanks to our methodology.

The investment and finance sector appears as the largest purveyor of cases of violations, representing 24% of total instances. Diversified banks alone represent 18% of the total; health and pharmaceuticals, ICT, and automobile each represent between 11% and 15%.

FIGURE 13

According to Clarity AI's methodology, 80% of companies with UNGC norms violations are located in Europe and North America.

Region of origin of companies violating UNGC norms¹



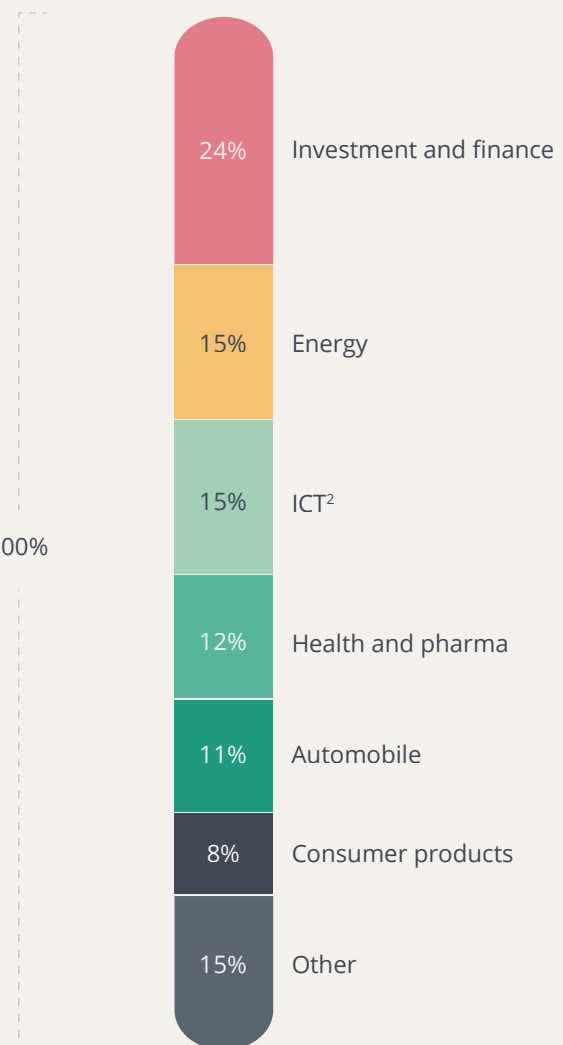
Note: Figures may not sum to 100%, because of rounding.
1. 100% = the total number of companies with biodiversity breaches.

The energy sector is the only sector accounting for a significant share of the principal adverse impacts for both biodiversity (49%) and UNGC and OECD principles violations (15%).

FIGURE 14

The investment and finance sector includes the most UNGC norms violations, with 24% of total instances.

Sector breakdown of companies violating UNGC norms¹



1. 100% = the total number of companies with biodiversity breaches.
2. Information and communications technology.

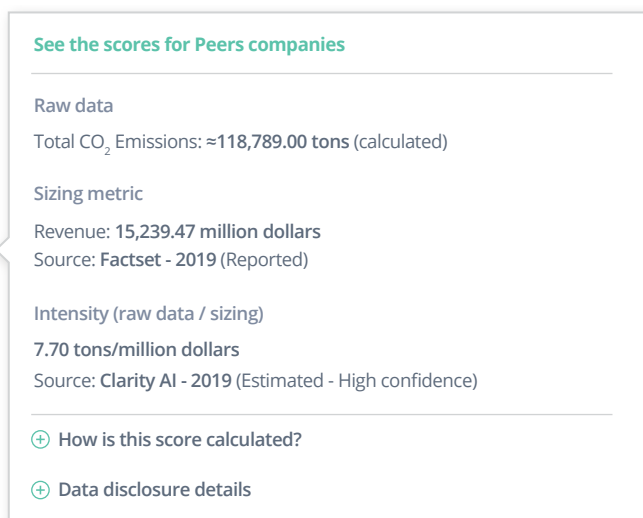


Clarity AI's commitment to transparency and continuous improvement

We view full transparency as a prerequisite for relying on data science to develop solutions for our clients. That is why Clarity AI indicates when data is reported versus when it is estimated, and when estimated, we include confidence flags to communicate the reliability of the estimates.

FIGURE 15

Given that transparency is a priority for Clarity AI, reported data includes indications of estimates and their reliability.



All estimates presented by Clarity AI undergo strict quality checks, taking into account the ability of the model to order companies correctly and to predict values with a distribution that is similar to the original distribution. In addition to clearing a quality-check threshold value, Clarity AI estimates are provided with a confidence level of low, medium, or high that reflects how well our estimates are expected to perform against real values in the holdout set. The confidence levels are set based on the value of the rank-order metric for the sub-industry associated with the company estimated.

Models are updated frequently as more reported data become available. Also of note is that machine learning's accuracy increases over time as the learning period increases. While we currently rely on more than 100 publishers, including NGOs, for our controversy model based on natural language processing, we are constantly looking to expand both our geographic and language coverage. This is the most promising way going forward to correct any potential geographic biases in the news media, as highlighted above.

As a next step, we will expand the coverage of our NLP-based controversy models to cover five additional voluntary PAI.

Conclusion

As greater transparency is required from investors by regulators and society at large on the characteristics and impact of their sustainable investments, the ability to provide sound sustainability performance metrics on the assets they invest in—mainly corporations—has become critical. While an unprecedented effort is underway to establish international standards for corporate sustainability reporting (CSRD in Europe and International Sustainability Standard Board), sustainability data will not be available at scale for at least another five to ten years.

Clarity AI's data science expertise, combining different techniques including machine learning and natural language processing, allows us to provide powerful solutions to overcome the current limitations of sustainability data. Through the combination of multiple data sources and data expertise, we have developed the most reliable dataset, outperforming any other single data provider. We then expanded that coverage by an average of 500% with estimation models. Through natural language processing, we developed solutions, characterizing adverse impacts on

biodiversity-sensitive areas and detecting UNGC and OECD violations with a high level of granularity for more than 16,000 companies.

However, we remain aware that data science also has its own limitations. We actively commit to reducing them by keeping humans in the loop with multidisciplinary teams, which bring together sustainability experts and data scientists. We fine-tuned our algorithm based on Facebook's RoBERTa model as a result of this expertise. While looking for controversy signals in news, we also implement mitigation measures to avoid bias in AI models, removing company names from the text so the algorithm neither learns them nor applies what it knows. Last, we avoid the aggregation of different models, optimizing robustness and simplicity.

Excellence is a core element of Clarity AI's values. Being fact-based and transparent is a strong marker of our culture. Our data science approach toward SFDR requirements brings these principles to life to deliver a unique solution to our customers with unprecedented reliability, robustness, and transparency.

Annex

FIGURE 16

Clarity AI groups controversial conducts into 39 categories.

Environmental	Social	Governance
Resource use Water use Animal well-being Land and bio impact	Employees Remuneration and working conditions Health and safety incidents	Corporate governance Corporate governance Accounting and taxation
Emissions Climate change mitigation Waste management impact	Customers and products Advertising and product representation incidents Customer data privacy incidents Media ethics incidents Product quality and safety responsibility Product social responsibility	Corporate ethics and behavior Business ethics Corruption incidents Human rights Inter-firm competition Sanctioned countries' and organizations' relations Suppliers business ethics Suppliers corruption incidents Weapons activity Intellectual property Lobbying
Suppliers footprint Suppliers' animal well-being impact Suppliers' climate change mitigation Suppliers' land and bio impact Suppliers' water stress Supply chain's waste management incidents	Supply chain Suppliers' human rights Suppliers' labor rights Suppliers' social impact Suppliers' health and safety	
Product footprint Product climate change impact Products environmental impact	Community and society Labor rights Population basic needs Local impact Suppliers local impact	



C L A R I T Y A I



Copyright

Copyright © 2021 Clarity AI all rights reserved.

Copyright © 2021 Clarity AI all rights reserved. The information contained in these documents is confidential, privileged and only for the information of the intended recipient and may not be used, published or redistributed without the prior written consent of Clarity AI.