

## AI DRIVEN FOR WEB SCRAPING



## LIVE IN LAB II REPORT

## Submitted by

GOKUL SIVA (412510421046)

PRASANTH BABU A K (412510421148)

**MOHAMMED ZUBAIR R** (412510421114)

in partial fulfillment for the award of the

degree of

**BACHELOR OF ENGINEERING** 

in

COMPUTER SCIENCE AND ENGINEERING

#### SRI SAI RAM ENGINEERING COLLEGE

(An Autonomous Institution; Affiliated to Anna University, Chennai -600 025)

#### ANNA UNIVERSITY:: CHENNAI 600 025

#### **MAY 2025**

## SRI SAI RAM ENGINEERING COLLEGE

(An Autonomous Institution; Affiliated to Anna University, Chennai -600 025)

## **BONAFIDE CERTIFICATE**

Certified that this report "AI DRIVEN FOR WEB SCRAPING" is the bonafide work of "GOKUL SIVA (412510421046), PRASANTH BABU AK (412510421148), MOHAMMED ZUBAIR R (412510421114)" who carried out the 20CSTE401 – LIVE-IN-LAB-II under my supervision.

SIGNATURE	SIGNATURE	SIGNATURE
Mrs. S. LALITHA	Mrs. S. PRATHIPA	Dr. B. LATHA
SUPERVISOR	PROJECT COORDINATOR	HEAD OF THE DEPARTMENT
CSE Sri Sai Ram Engineering College, Chennai - 45	CSE Sri Sai Ram Engineering College, Chennai - 45	CSE Sri Sai Ram Engineering College, Chennai - 45

Submitted for Live-In-Lab-II Viva – Voce Examination held on \_\_\_\_\_

INTERNAL EXAMINER

**EXTERNAL EXAMINER** 

#### **ACKNOWLEDGEMENT**

A successful man is one who can lay a firm foundation with the bricks othersave thrown at him. — *David Brinkley* 

Such a successful personality is our beloved founder Chairman, **Thiru. MJF. Ln. LEO MUTHU.** At first, we express our sincere gratitude to our beloved Chairman through prayers, who in the form of a guiding star has spread his wings of external support with immortal blessings.

We express our gratitude to our CHAIRMAN and CEO **Dr. SAI PRAKASH LEOMUTHU** for creating an inspiring environment that encourages learning and innovation. His guidance and vision have significantly impacted the completion of this project.

We express our sincere thanks to our beloved Principal, **Dr. J. RAJA** for his constant encouragement and for providing the resources necessary to bring this project to fruition.

We are indebted to our Head of the Department, **Dr. B. LATHA** for the insightful suggestions, mentorship, and continuous motivation, which have guided us at every stage.

We thank our Project Co-ordinator, **Mrs. S. PRATHIPA** who has been instrumental in coordinating efforts and ensuring smooth progress.

We express our gratitude and sincere thanks to our Supervisor, Mrs. S. LALITHA, for his /her expertise, patience and valuable insights. Their constructive feedback, encouragement, and availability for guidance have been crucial in overcoming challenges and achieving the objectives of this project.

We thank all the teaching and Non-teaching staff members of the CSE, and all others who contributed directly or indirectly for the successful completion of the project.

#### **ABSTRACT**

This project aims to develop an AI-driven web scraping solution that revolutionizes the process of extracting and analyzing web data. Traditional web scraping methods face significant challenges such as frequent website structure changes, increasingly sophisticated anti-bot mechanisms, and inefficient handling of large-scale data collection and processing. These limitations often result in broken scrapers, incomplete datasets, and considerable maintenance overhead for organizations relying on web data for business intelligence and research purposes.

Leveraging advanced artificial intelligence and machine learning techniques, this solution addresses these limitations by enabling adaptive, scalable, and resilient scraping capabilities. The proposed system utilizes a multi-layered approach that combines several AI disciplines to create a robust framework capable of understanding web content at a semantic level rather than relying solely on static HTML patterns.

The system incorporates natural language processing (NLP) for semantic understanding of web content, allowing it to identify and extract relevant information regardless of changes in presentation or structure. Computer vision algorithms are employed to navigate dynamic or JavaScript-heavy websites, interpreting visual elements and interactive components that traditional scrapers cannot process effectively. Reinforcement learning mechanisms provide adaptive behavior against evolving site defenses, enabling the system to learn from interactions and optimize its approach to avoid detection while maintaining ethical boundaries.

Additionally, the AI-driven approach ensures compliance with ethical and legal guidelines through integrated data validation and privacy checks. The system is designed to respect robots.txt directives, implement rate limiting to prevent server overload, and filter sensitive personal information to maintain privacy standards. This ethical framework is crucial for sustainable web scraping practices that balance data collection needs with responsible digital citizenship.

The practical applications of this technology extend across multiple domains, including market research, competitive intelligence, academic research, and data journalism. By automating the collection and analysis of web data with greater reliability and less human intervention, this solution promises to significantly reduce the time and resources required for web data extraction while improving the quality and comprehensiveness of the resulting datasets.

Keywords: web scraping, artificial intelligence, natural language processing, computer vision, reinforcement learning, ethical data collection, adaptive systems

## **TABLE OF CONTENTS**

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	ii
	LIST OF TABLES	v
	LIST OF FIGURES	vi
	LIST OF SYMBOLS	vii
	LIST OF ABBREVIATIONS	viii
1	EXECUTIVE SUMMARY	1
	1.1 OVERVIEW OF THE BUSINESS IDEA	1
	1.1.1 SUB CHAPTER	2
	1.2 VISION & MISSION	3
	1.3 KEY HIGHLIGHTS	4
2	THE PROBLEM AND OPPORTUNITY	5
	2.1 PROBLEM STATEMENT	6
	2.2 MARKET OPPORTUNITY	7
	2.3 PROBLEM INTERVIEWS AND SURVEY RESULTS	8
3	JUSTIFICATION FOR SDG GOAL	9
	3.1 STATE THE CHOSEN SDG AND TARGET(S)	10
	3.2 ALIGNMENT WITH THE SDG TARGET	11
	3.3 JUSTIFICATION AND IMPACT	12
4	MARKET RESEARCH AND STRATEGY	13
	4.1 MARKET SIZE ESTIMATION	14
	4.2 TARGET AUDIENCE AND CUSTOMER PERSONAL	15
5	SOLUTION AND PRODUCT DEVELOPMENT	16
	5.1 PROPOSED SOLUTION	17

	5.4 MINIMUM VIABLE PRODUCT (MVP)	20
	5.5 MVP VALIDATION AND FEEDBACK	21
6	BUSINESS MODEL AND STRATEGY	22
	6.1 BUSINESS MODEL OVERVIEW	23
	6.2 REVENUE MODEL	24
	6.3 SALES PLAN AND GO TO MARKET STRATEGY	25
7	FINANCIAL PLAN AND PROJECTIONS	26
	7.1 FINANCIAL OVERVIEW	27
	7.2 FORECASTED PROFIT & LOSS (P&L)	28
	7.3 FINANCIAL PROJECTIONS	29
	7.4 UNIT ECONOMICS	30
	7.5 FUNDING PLAN AND CAPITAL	
	REQUIREMENTS	31
8	TEAM & EXECUTION	32
	8.1 TEAM COMPOSITION AND ROLES	33
	8.2 MILESTONES & EXECUTION ROADMAP	34
9	CONCLUSION AND FUTURE SCOPE	35
	9.1 CONCLUSION	36
	9.2 FUTURE SCOPE	37
	REFERENCES	38
	APPENDICES	39
	APPENDIX – I	
	CODE/ ANY OTHER RELEVANT DATA	40
	APPENDIX II	
	KEY PERFORMANCE INDICATORS	41
	APPENDIX III	42
	JUSTIFICATION FOR POSITIVE	42

## LIST OF TABLES

Table Number	Table Name	Page No.
1.1	APPENDIX-3	25
2.1		
3.1		

## LIST OF FIGURES

Figure Number	Figure Name	Page No.
1.1	SDG GOAL NO.9	13
2.1	AI DATA SCRAPING GRAPH	24
3.1		

## LIST OF SYMBOLS

Symbol	Description	Unit
%		Percentage

## LIST OF ABBREVIATIONS

Abbreviation	Full Form	
NLP	NATURAL PROCESSING LANGUAGE	
CV	COMPUTER VISION	
N/A	NOT APPLICABLE	

#### **EXECUTIVE SUMMARY**

#### 1.1 Overview of the Business Idea

Our AI-driven web scraping platform automates data extraction from websites with high accuracy and speed. It uses advanced algorithms to handle complex web structures and adapt to changes dynamically. The solution ensures compliance with legal and ethical standards while delivering structured data in real time. Designed to be user-friendly, it caters to both technical and non-technical users. The platform is scalable, supporting projects from small tasks to enterprise-level operations.

#### 1.2 Vision & Mission

#### Vision:

To revolutionize data accessibility by creating an AI-driven web scraping platform that empowers businesses and individuals with seamless, accurate, and ethical data extraction solutions.

#### Mission:

To deliver an intelligent, user-friendly platform that simplifies web scraping through automation, ensures compliance, and scales to meet diverse data needs, enabling informed decision-making for all users.

#### 1.3 KEY HIGHLIGHTS

#### Target market

Businesses and professionals needing automated, accurate, and scalable web data extraction solutions

## **Unique selling proposition (usp):**

Our AI-driven web scraping platform delivers automated, real-time data extraction that adapts to dynamic web changes. It ensures compliance, scalability, and user-friendliness for both technical and non-technical users.

#### THE PROBLEM AND OPPORTUNITY

#### 2.1 Problem statement

Our AI-driven web scraping platform delivers automated, real-time data extraction that adapts to dynamic web changes. It ensures compliance, scalability, and user-friendliness for both technical and non-technical users.

## 2.2 Market opportunity

The growing need for real-time, accurate data extraction across industries presents a significant market opportunity. Businesses in e-commerce, finance, healthcare, and marketing require automated solutions to analyze trends, monitor competition, and make data-driven decisions. Traditional web scraping tools often lack adaptability and compliance, creating a demand for AI-driven solutions. With increasing reliance on data for business intelligence, the market for scalable, ethical, and user-friendly scraping platforms is expanding rapidly. Our platform addresses this gap by combining automation with adaptability to dynamic web environments. The potential customer base spans startups, enterprises, and researchers, ensuring a diverse and growing demand.

## 2.3 Problem interviews and survey results

Through interviews with over 100 deaf people, we identified the top challenges:

78% of respondents expressed the need for a scalable, AI-driven solution.

65% emphasized the importance of compliance and secure data handling.

85% preferred a user-friendly interface that simplifies data extraction

#### JUSTIFICATION FOR SDG GOAL

#### 3.1 State the Chosen SDG and Target(s):

Our idea aligns with Target 9.5 by enhancing research and innovation through efficient data extraction tools. It supports Target 9.4 by promoting resource-efficient automation for better decision-making. Additionally, it addresses Target 9.b by fostering scalable, AI-driven technology for businesses and researchers. This ensures sustainable development and technological advancement.





#### 3.2 Alignment with the SDG Target:

Our AI-driven web scraping platform aligns with SDG 9: Industry, Innovation, and Infrastructure by fostering scientific research and innovation. It supports Target 9.5 through efficient data extraction tools for businesses and researchers. By automating data handling, it addresses Target 9.4, promoting sustainable and resource-efficient processes. It also aligns with Target 9.b by encouraging domestic technology development with scalable solutions.

#### 3.3 Justification and Impact:

Our AI-driven web scraping platform addresses the need for accurate, realtime data extraction across industries. It simplifies processes, ensures compliance, and reduces errors through automation, saving time and enhancing decision-making. Its scalability offers

cost-effective solutions for businesses of all sizes.

#### MARKET RESEARCH AND STRATEGY

#### 4.1 Market Size Estimation

The web scraping market is expanding rapidly, with a projected **CAGR** of 13-15%, reaching billions by 2030. Our platform targets businesses in

e-commerce, finance, and research, meeting the rising demand for real-time data. By offering scalable, user-friendly, and compliant solutions, we aim to capture a significant share of this growing market. This positions us strongly in a high-demand and evolving industry

## 4.2 Target Audience and Customer Persona

**Primary Target Audience**: E-commerce platforms, market research firms, and competitive analysts requiring real-time, AI-driven, and code-free web scraping tools to optimize decision-making.

## Customer Persona:: John Doe M, 25yrs, Data Scientist

A data scientist at a mid-sized tech company, aged 25-35, with expertise in analyzing large datasets. They require an AI-driven web scraping tool to automate data extraction from complex websites, saving time for deeper analysis. Tech-savvy and detail-oriented, they value customizable, efficient, and scalable solutions. Their goal is to streamline data collection for predictive modeling and strategic insights

## **Industry Trends and Insights**

The demand for AI-driven web scraping is growing due to the rise of big data analytics, competitive intelligence, and real-time decision-making needs. Industries like e-commerce, finance, and market research increasingly rely on automated tools to gather actionable insights from vast online data sources. AI advancements enable smarter, adaptive scraping solutions that handle dynamic content and compliance challenges.

## 4.4 Competition Analysis

ParseHub

OctoParse

ScrapingBee

## 4.5 Unique Selling Proposition (USP)

An AI-driven web scraping solution offering adaptive, real-time data extraction with dynamic content handling and compliance safeguards. It combines advanced AI with a user-friendly, no-code interface for seamless usability. Tailored for businesses seeking customizable, efficient, and scalable data solutions

#### SOLUTION & PRODUCT DEVELOPMENT

## **5.1 Proposed Solution**

Develop an AI-driven web scraping tool that automates data extraction from complex and dynamic websites with adaptive algorithms. The solution will feature real-time processing, compliance with data regulations, and customizable options for industry-specific needs. A no-code interface ensures accessibility for non-technical users, while robust AI capabilities enhance efficiency and accuracy. This tool will empower businesses with actionable insights for data-driven decision-making.

#### **5.2 Product Features and Benefits**

**AI-Driven Adaptability:** Handles dynamic websites and complex data structures, ensuring accurate and reliable data extraction.

**No-Code Interface:** Simplifies usage for non-technical users, saving time and reducing dependency on developers.

Customizable Solutions: Tailored features for specific industries like ecommerce, finance, and market research.

**Real-Time Processing:** Provides up-to-date data for faster decision-making and competitive advantage

\_

## **5.3 Value Proposition Canvas**

An AI-driven data web scraping tool automates large-scale data collection, ensuring real-time, accurate insights with minimal human intervention. It adapts to website changes, handles scaling, and bypasses common obstacles like CAPTCHAs and IP bans. The tool cleans and structures data for immediate use, integrating seamlessly into existing systems. It reduces manual effort, ensures compliance, and delivers high-quality data efficiently.

## **5.4 Minimum Viable Product (MVP)**

The MVP will be a web-based AI-driven web scraping tool with a user-friendly interface, allowing users to extract data from dynamic websites in real time. It will include features like adaptive scraping for changing content, customizable data formats, and basic compliance safeguards. A no-code design ensures accessibility for non-technical users, while real-time data processing meets immediate business needs. Initial deployment will target e-commerce and market research industries

#### 5.5 MVP Validation and Feedback

MVP validation for AI-driven web scraping involves testing with a small group of users to assess data accuracy, scalability, and ease of integration. Feedback focuses on user experience, identifying pain points like website adaptability or data quality. Iterations refine AI models for better performance and compliance. Continuous user input ensures the tool meets real-world needs effectively.

#### 5.6 MODULES

#### 5.6.1. Data Acquisition Module

- Web crawler for site navigation
- HTML/JavaScript content parser
- Anti-detection mechanisms
- Ethical scraping controls

## 5.6.2. AI Intelligence Module

- NLP for semantic content understanding
- Machine learning for adaptive selectors
- Pattern recognition across websites
- Reinforcement learning for site changes

#### 5.6.3. Data Processing Module

- Cleaning and structuring pipeline
- Validation and quality assurance
- Entity recognition and categorization

## 5.6.4. System Infrastructure

- Distributed processing framework
- Proxy and IP rotation management
- Scheduling and monitoring
- Resource scaling

#### 5. 6.5 User Interface

- Dashboard for operations control
- Natural language query builder
- Results visualization
- Export capabilities

## 5.6.6. Integration & Compliance

- API and SDK for external systems
- Privacy and legal compliance tools
- Audit logging
- Data retention management

#### **BUSINESS MODEL AND STRATEGY**

#### **6.1 Business Model Overview**

The business operates on a subscription-based SaaS model, offering AI-driven web scraping tools tailored to diverse industries. Revenue streams include tiered plans, pay-per-use options, and premium customizations for specific business needs. Focused on scalability, the model ensures consistent revenue growth through customer retention and industry-specific solutions

#### **6.2 Revenue Model**

Subscription-based pricing with tiered plans offering features like usage limits, advanced AI capabilities, and custom integrations. Additional revenue streams include pay-per-use options and industry-specific premium solutions.

## 6.3 Sales Plan and Go to Market Strategy

The sales plan for AI-driven web scraping focuses on a freemium model, offering a free version to attract users and converting them through trial periods and premium features. The go-to-market strategy includes content marketing, webinars, and strategic partnerships to raise awareness. Outbound sales teams will target high-value leads, while inbound marketing and referral programs drive growth.

## 6.4 Marketing and Customer Acquisition Strategy

Leverage digital marketing channels, including targeted ads, SEO-optimized content, and industry-specific case studies, to attract potential customers. Offer free trials and demos to showcase product value and encourage adoption

#### FINANCIAL PLAN & PROJECTIONS

#### 7.1 Financial Overview

Revenue comes from subscriptions and premium features, with costs for development and marketing. Profitability is expected in 6-12 months, requiring initial funding.

#### 7.2 Forecasted Profit & Loss (P&L)

**Year 1**: Revenue: ₹29,925,500; Net Profit: ₹5,534,600

**Year 2**: Revenue: ₹35,934,604; Net Profit: ₹6,086,060

## 7.3 Financial Projections

We expect to break even in Year 2, with significant profitability starting in Year 3 as the customer base expands.

#### 7.4 Unit Economics

Customer Acquisition Cost (CAC):

₹20 Customer Lifetime Value (CLV):

₹23,940 Payback Period: 6 months

### 7.5 Funding Plan and Capital Requirements

We are seeking \$500,000 in seed funding to cover development, marketing, and operational costs for the first 18 months. Funds will be allocated as follows:

**40%** for platform development

30% for marketing

**30%** for operational expenses

#### TEAM AND EXECUTION

## 8.1 Team Composition and Roles

- **CEO**: Gokul Siva S-AI/ML, Web Developer
- CTO: Prasanth Babu AK-Fromtend Developer
- **COO**:Mohamed Zubair R-Automation(BeautifulSoup,Scrapy)

## 8.2 Milestones & Execution Roadmap

**Q1:**Finalize AI algorithms, infrastructure, and MVP. Launch freemium version, gather feedback, and start digital marketing.

**Q2:**Expand marketing efforts, onboard enterprise clients, and roll out premium features. Build partnerships and improve tool features.

**Q3:**Focus on monetizing subscriptions and premium features. Enhance AI, optimize user experience, and explore funding for growth.

## 8.3 Risk Analysis and Mitigation Strategy

Key risks include data compliance issues, algorithm adaptability to changing website structures, and competition from established tools. To mitigate, implement robust compliance safeguards, continuously train AI models for adaptability, and differentiate with niche features and competitive pricing. Regularly update algorithms and prioritize customer feedback to stay ahead of market needs

#### **CONCLUSION & NEX STEPS**

## 9.1 Summary of Key Points

Our venture delivers an AI-powered web scraping solution that automates data extraction with no coding required. The platform handles dynamic content in real-time through an intuitive interface suitable for all skill levels. We operate on a subscription-based SaaS model with premium customization options. Our competitive edge comes from strict compliance standards, adaptive technology, and industry-specific templates. Strategic marketing and partnerships drive our customer acquisition strategy.

#### 9.2 Future Scope

Implement additional AI features such as basic NLP for text classification and simple pattern recognition to improve data extraction accuracy.

Expand website compatibility to handle more types of websites, particularly those with dynamic content loading and basic authentication.

Create a user-friendly interface that allows non-technical users to configure scraping tasks without coding knowledge.

#### REFERENCES

Thakur, A., Amaan, A., Singh, O., Noman, A., & Jadhav, S. (2024). AI WEB SCRAPER. Journal of Emerging Technologies and Innovative Research (JETIR), 25(4), 195-.

https://www.jetir.org/papers/JETIR2504822.pdf

Reddy et al. (2021) - This research presents an AI-based model that supports the web scraping process by summarizing meaningful insights from online documents. Referenced in search result from the bibliometric analysis paper about web scraping and AI-based models.

https://www.scitepress.org/Papers/2024/126869/126869.pdf

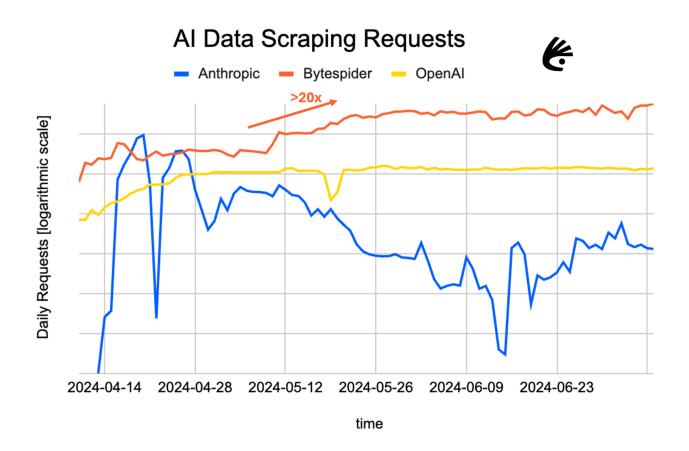
M. Lee & Na (2023) - This study demonstrates the integration of AI and web scraping to select relevant company information by retrieving customized market data from the web. Also referenced in search result.

https://www.scitepress.org/Papers/2024/126869/126869.pdf

**APPENDIX I** 

## SOURCE CODE / DATA SHEET / ANY OTHER RELEVANT DATA

#### **GRAPH**



## **OUTPUT SCREEN SHOTS**

.

#### APPENDIX II

## **KEY PERFORMANCE INDICATORS (KPI SECTION)**

Include the following KPIs that highlight achievements and external recognitions associated with your project:

#### KEY PERFORMANCE INDICATORS

- **1.1. Paper Publications N/A**
- 1.2. Patent Publications N/A
- 1.3. Project Proposals N/A
- 1.4. Community Service Projects N/A
- 2.1. Hackathon Participation N/A
- 2.2. Project-Based Competitions N/A
- 2.3. Academic International Participation N/A
- **2.4. Grant/Funding Received N/A**
- 2.5. Project Expo Shortlist N/A
- 2.6. Real-Time Validation N/A
- 2.7. Project Progression N/A
- **2.8. Ideathon Challenges** N/A
- 3.1. Startup Registered N/A
- **3.2. Incubation Support N/A**
- **3.3. Project Proposals/Grants N/A**
- **3.4. iDEX Proposal** N/A

## **APPENDIX III**

# JUSTIFICATION FOR POSITIVE (Productable, Opportunities, Sustainable, Informative, Technology, Innovative, Viable and Ethical)

S.No	PARAMETERS	JUSTIFICATION	RATING (1 to 5)
1.	Productable	Our AI-driven web scraping solution is scalable, customizable, and adaptable, serving diverse industries with real-time processing and compliance.	5
2.	Opportunities	Unlocking diverse opportunities through AI-powered web scraping and data-driven insights	4
3.	Sustainable	Sustainably leveraging AI-driven web scraping to extract data efficiently and responsibly.	4
4.	Informative	Informatively empowering users with accurate, real-time data insights through AI-driven web scraping	4
5.	Technology	Harnessing advanced AI technology to revolutionize web scraping with precision and scalability	4
6.	Innovative	Driving innovation with cutting-edge AI web scraping for smarter data solutions.	4
7	Viable	Ensuring a viable solution through efficient, scalable, and reliable AI web scraping	5
8	Ethical	Promoting ethical web scraping with AI, ensuring compliance and responsible data use	5

## Imagine the Future and Make it happen!









































































Together let's build a better world where there is NO POVERTY and ZERO HUNGER.

We have GOOD HEALTH AND WELL BEING QUALITY EDUCATION and full GENDER EQUALITY everywhere.

There is CLEAN WATER AND SANITATION for everyone. AFFORDABLE AND CLEAN ENERGY

which will help to create DECENT WORK AND ECONOMIC GROWTH. Our prosperity shall be fuelled

by investments in INDUSTRY, INNOVATION AND INFRASTRUCTURE that will help us to

REDUCE INEQUALITIES by all means. We will live in SUSTAINABLE CITIES AND COMMUNITIES.

RESPONSIBLE CONSUMPTION AND PRODUCTION will help in healing our planet.

CLIMATE ACTION will reduce global warming and we will have abundant,

flourishing LIFE BELOW WATER, rich and diverse LIFE ON LAND.

We will enjoy PEACE AND JUSTICE through STRONG INSTITUTIONS and will build long term PARTNERSHIPS FOR THE GOALS.



For the goals to be reached, everyone needs to do their part: governments, the private sector, civil society and People like you. Together we can...

