Exploring Safe, Secure, and Reliable (SSR) AI

Module Overview:

This module introduces learners to the fundamental concepts of safe, secure, and reliable AI (SSR AI) and its importance in the field of artificial intelligence. It covers the definition of AI, machine learning (ML), and deep learning (DL) in the context of SSR AI. Additionally, the module dives into the types of problems that are solvable to AI solutions and how AI formulates solutions to these problems. It also discusses various branches and capabilities of AI beyond machine learning, such as reasoning, planning, and knowledge representation, and introduces related fields within computational intelligence.

Learning Objectives:

By the end of this module, students should be able to:

1. Define AI, ML, and DL in the context of SSR AI.
2. Explain the importance of SSR AI in ensuring trust and fairness in AI systems.
3. Identify types of problems that can be addressed using AI solutions.
4. Formulate problems for potential AI solutions.
5. Explore branches and capabilities of AI beyond machine learning.
6. Recognize synergistic fields in computational intelligence.

Outline:

1. Introduction to Artificial Intelligence, Machine Learning, Deep Learning in the context of SSR AI
2. Types of Problems that can be solved by using Artificial Intelligence
3. How does AI formulate a solution to a problem?
4. Branches and Capabilities of an AI
5. Memetic Computing and Genetic Computing

Terminology:

- **Algorithm**: An algorithm is a set of step-by-step instructions that tell a computer how to do something.

- **Bias:** Bias is when a computer makes unfair decisions because of the data it was trained on.
- **Big Data:** Big data is data that is so big it humans are not capable of processing.
- **Data**: Data is information, like numbers, words, or images, that computers use to learn and make decisions.
- **Natural Language Processing (NLP):** NLP is how computers understand and talk with humans using language, like when you chat with a virtual assistant.
- **Neural Network:** A neural network is a model inspired by the human brain. It helps computers understand and recognize complex patterns.
- **Prediction:** A prediction is a guess a computer makes based on what it has learned.
- **Training Data:** Training data is the information used to teach a computer. It's like giving examples for a computer to learn from.

Lesson 1: Introduction to Safe, Secure and Reliable Artificial Intelligence, Machine Learning, Deep Learning

1) Introduction to Safe, Secure, Reliable (SSR) AI

There are various definitions for Artificial Intelligence, and it changes from context to context. How does AI define itself? ChatGPT defines AI as the "simulation of human intelligence in machines or computer systems". On the other hand, computer scientist, John McCarthy defines AI as "It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable" in his article "What is Artificial Intelligence?". Both definitions are correct. In simple terms, AI is the imitation of the human brain.

What are the applications of artificial intelligence? AI is everywhere in our daily lives. For instance, the smartphones voice assistants like Siri, Google Assistant, and Bixby use AI to understand and respond to spoken commands. Another example can be on social media AI algorithms power content recommendations, friend suggestions, and content moderation, helping personalize the content the user is seeing. AI also filters out spam emails and categorizes user's inbox, making it easier to manage and prioritize messages. Is technology that is this useful and part of our daily lives safe, secure, and reliable? We are going to discuss this in this lesson.

a. What makes AI secure, safe, and reliable?

AI is an imitation of human intelligence. However, the key difference is that humans have values, and they make their decisions based on that. On the other hand, AI's decisions are based on calculations. The safety and security of the AI's is up to the creators of the AIs. What are the key problems of AI safety, security, and reliability?

i. **Secure:** AI is secure if it is resistant to attack and unauthorized access.

AI systems are often used to process and store sensitive data, such as financial information, medical records, and personal data.

ii. **Safe:** AI safety is very important when building an AI. AI is considered safe if it does not cause harm to humans or the environment. AI systems often make decisions that have real life consequences.

AI systems are used to control self-driving cars, diagnose diseases, and recommend financial investments. If an AI system is not safe, it could make decisions that lead to accidents, injuries, or financial losses.

iii. **Reliable:** AI can be considered reliable when it gives consistent and accurate results.

AI systems are used to make important decisions, for instance to approve a loan or grant a medical license.

b. What are the barriers of AI to be safe, secure, and reliable?

1) Clear Instructions

When training AI, it is important to give it a clear instruction of what we want it to do. Lack of clear instructions can lead AI to give out results that user didn't mean to ask. For instance, when we are training the AI model to recognize a fruit image, we need to make sure that we include all the fruits. Not just summer fruits, or fruits grow only in the US. AI can pick up a pattern quickly and turn this into a summer fruit that grows in the US.

2) Transparency

AI follows a certain pattern when it's deciding. However, sometimes it can be hard for people who wrote the algorithm to know why AI chooses the pattern that it's choosing. Caltech professor, Yaser Abu-Mostafa says that "Scientifically, we don't know why the neural networks are working as well as they are. If you look at the math, the data that the neural network is exposed to, from which it learns, is insufficient for the level of performance it attains." For AI to be used in real-world decision-making problems, we need to make sure that we know why AI follows a certain pattern. Otherwise, it can lead to a bigger issue in the future.

3) Uncertainty Measures

AIs tend to make mistakes just like humans. However, they tend to be overconfident, and this might mislead the user to realize that AI is making a mistake. AI needs to determine these mistakes before giving an answer. Designing this part of an AI is a technical challenge.

4) Bias in data

The data that is given to train the AI is very important. The larger data set is better for the training of the AI. However, it can be hard to control the dataset if it is too large. This can lead to some problems, such as bias in the AI.

2) What is Machine Learning (ML)?

Machine learning is a branch of AI that focuses on developing algorithms that can learn from data and improve their performance over time without being explicitly programmed.

a. Safe, Secure and Reliable Machine Learning

Machine Learning techniques must be employed carefully, ensuring the data used for training and the algorithms applied are robust, unbiased, and capable of producing valid and reliable results to support trust and fairness.

Additionally, there can be attacks during the development and training process of machine learning. Attackers can try to “poison” the data that ML is training with. This can cause ML models to make poor decisions. For instance, when self-driving cars are poisoned during the training process, it can cause car crashes, etc.

3) What is Deep Learning (DL)?

Deep learning is a branch of machine learning, and it works with neural networks that have three or more layers. These networks try to imitate how our brains work. They are great at learning from a lot of information. If you have just one layer in the network, you can make rough guesses, but if you add more hidden layers, it gets better and more precise.

a. Safe, Secure and Reliable Deep Learning

Deep Learning models can be hard to understand. It can be difficult to understand how complex deep learning models make decisions, and it can be difficult to ensure that models are unbiased, especially when they are trained on large and diverse datasets. It also is important to prioritize security, and fairness. This involves implementing measures to protect against cyber threats, making models more interpretable, and ensuring unbiased outcomes.

Activities:

1. Group Discussion: Real-world Applications of AI
   Objective: To understand the practical applications of AI and identify where SSR AI principles are crucial.
   Activity: Divide the class into small groups. Each group discusses different AI applications in daily life (like voice assistants, recommendation systems, self-driving cars). They need to identify potential safety, security, and reliability concerns for each application.
2. Case Study Analysis: AI Failures
   Objective: Analyze real-life scenarios where AI systems failed due to lack of safety, security, or reliability.
   Activity: Provide students with case studies of AI failures (like biased decision-making, security breaches in AI systems). Students will analyze the root causes of these failures and propose solutions or preventive measures.
3. Role Play: AI Ethics Committee

Objective: To understand the importance of ethical considerations in AI development.
Activity: Students role-play as members of an AI ethics committee. They evaluate a proposed AI project, considering aspects of safety, security, and reliability, and make a decision whether to approve, reject, or suggest modifications to the project.

4. Designing a Secure AI System
   Objective: Apply knowledge of AI security in a practical setting.
   Activity: Students design a basic outline of a secure AI system. They must include elements that ensure data protection, resistance to cyber-attacks, and safe decision-making processes.
5. Debate: Bias in AI
   Objective: To encourage critical thinking about bias in AI and its implications.
   Activity: Organize a debate on the statement "All AI systems inherently carry the bias of their creators." Students prepare arguments for and against, focusing on how biases affect AI safety and reliability.
6. Interactive Quiz
   Objective: Test knowledge of key terms and concepts related to SSR AI.
   Activity: Create an interactive quiz with questions about AI, ML, DL, neural networks, bias, etc. This can be done using digital tools to make it more engaging.
7. Predicting AI Trends
   Objective: To encourage forward-thinking about the future of AI.
   Activity: Students write a short essay predicting the future of AI in terms of safety, security, and reliability. They should base their predictions on current trends and their understanding of the field.
8. Reflection Journal
   Objective: Encourage self-reflection on learning.
   Activity: Students maintain a journal throughout the course, reflecting on what they learned after each lesson. For this lesson, they focus on the importance of SSR in AI.

Lesson 2: Problems that can be solved using AI, how does AI generate these solutions?

a. **Different problems that can be solved using AI:**

1) **Health**
   AI is an important tool that can be used in the healthcare sector. AI technology can help the doctors and the patients in many areas. Some examples of AI use in healthcare sector include, scanning surgical tools before the surgery and after the surgery to determine if any of them are left in the patient. Additional to this, doctor and patient time is very limited, and a doctor cannot know everything in the medical literature. AI can step up in this case and help the doctor with the diagnosis of the patient.
2) **Transportation**
   Self-driving cars are already a part of our world. The algorithms that are designed to make a car drive itself and make decisions based on the real-world issues. AI can also predict when vehicles need maintenance, helping traffic move better, reducing crowded roads. For instance, when there is a traffic and maps give you an alternative road, that's one of the instances where we use these algorithms.
3) **Education**
   AI can be adapted to each student's individual need. AI can be used as a tutor, where it explains the classroom concepts to the student. AI can be a good tool to give feedback to students so they will know their weaknesses. It can also be useful for grading and can help teachers save some time.

4) **Energy**
   International Energy Agency mentions that "One of the most common uses for AI by the energy sector has been to improve predictions of supply and demand. Developing a greater understanding of both when renewable power is available and when it's needed is crucial for next-generation power systems."

b. **Steps AI follows to generate a solution to the problem:**

1) **Problem Definition:** AI starts by identifying the problem it needs to solve. AI needs to understand the task, its objectives, constraints, and the data or information available for solving the problem. Clear problem definition is important for the success of AI solutions.

2) **Data Collection and Preprocessing:** AI relies heavily on data to make decisions. Data should be collected, cleaned, and preprocessed to ensure that it is of high quality, relevant, and suitable for analysis.

3) **Feature Selection/Extraction:** In many AI applications, relevant features or attributes are selected from the dataset or new features are extracted to represent the

problem more effectively. This helps in reducing dimensionality and improving the quality of input data.

4) **Algorithm/Model Selection:** AI uses a specific algorithm or model to solve the problem. The choice of algorithm depends on the nature of the problem, the available data, and the desired outcome. Common AI algorithms include machine learning models (e.g., neural networks, decision trees, support vector machines) and optimization algorithms.

5) **Model Training:** If machine learning is involved, the selected model is trained using the preprocessed data. During training, the model learns from the data to make predictions or decisions. This process often involves adjusting model parameters to minimize errors or optimize a specific objective function.

6) **Model Evaluation:** After training, the AI model is evaluated using a separate dataset (e.g., a validation or test set) to assess its performance.

7) **Hyperparameter Tuning:** Hyperparameters are parameters that are not learned during training but need to be set beforehand. AI systems often undergo hyperparameter tuning to find the best configuration for the model, maximizing its performance.

8) **Deployment:** Once the model is trained and evaluated, it can be deployed in a production environment to provide solutions to the problem. Deployment may involve integrating the AI system into a software application or a larger ecosystem.

Activities

1. Group Project: AI Solutions in Different Sectors
   Objective: To explore and understand the application of AI in various sectors like health, transportation, education, and energy.
   Activity: Students are divided into groups, each assigned to one sector (e.g., health, transportation). Each group research how AI is used in their sector, focusing on specific problems AI solves, and presents their findings to the class.
2. Role Play: AI Problem-Solving Process
   Objective: To understand the process of how AI generates solutions to problems.
   Activity: Students role-play the steps of AI problem-solving (from problem definition to deployment). Each student or group of students takes a step and explains its significance in the process, perhaps using a common problem as an example.
3. Data Collection and Analysis Exercise
   Objective: To gain practical experience in the initial steps of AI problem-solving.
   Activity: Provide students with raw data (or have them collect it). They then clean, preprocess, and analyze this data, mimicking the early steps of the AI problem-solving process.
4. AI Algorithm Selection Challenge

Objective: To learn about different AI algorithms and their applications.
Activity: Present students with several hypothetical problems. They must research and decide which AI algorithms would be best suited to solve each problem and explain their choices.

5. Machine Learning Model Training Workshop
   Objective: To gain hands-on experience in training a basic machine learning model.
   Activity: Using simple tools like Google's Teachable Machine or Python libraries (for more advanced students), have students train a basic model. This could involve image recognition, text classification, etc.
6. Model Evaluation Lab
   Objective: Understand the importance of evaluating AI models.
   Activity: Provide students with a pre-trained AI model and a dataset for testing. Students evaluate the model's performance and suggest improvements.
7. Hyperparameter Tuning Interactive Session
   Objective: To learn about the importance of hyperparameters in AI models.
   Activity: Using a simple machine learning model, students experiment with different hyperparameters to see how they affect the model's performance. This can be done through simulations or using AI learning platforms.
8. AI Ethics Discussion: Bias in Data
   Objective: To encourage critical thinking about ethical considerations in AI.
   Activity: Host a class discussion or debate on how bias in data collection can impact AI solutions, especially in sensitive areas like healthcare or law enforcement.
9. Case Study Analysis: AI Deployment
   Objective: Analyze the challenges and considerations in deploying AI solutions.
   Activity: Students review case studies of AI deployment in different sectors, discussing the challenges faced and the strategies used to overcome them.

Lesson 3: Branches and Capabilities of Artificial Intelligence

1) **Natural Language Processing (NLP):** It gives AI the ability to understand speech and text the same way human beings can.
    a. **Speech Recognition:** Converts spoken words to text.
    b. **Part of speech tagging (grammatical tagging):** Determines what context the word is used.
    c. **Word sense disambiguation:** Determines the word that might have multiple meanings. For instance, 'make the grade' (achieve) vs. 'make a bet' (place).
    d. **Name entity recognition (NEM):** It identifies 'Mary' as a name or 'Michigan' as a location.
    e. **Co-reference resolution:** It identifies 'she' as 'Mary', or an idiom in the text.
    f. **Sentiment analysis:** Tries to extract attitudes, emotions, sarcasm, confusion, suspicion from text.
    g. **Natural language Generation:** Putting structured information into human language.
2) **Computer Vision:**
    a. **Image Recognition:** Can classify an object after seeing it (a cat, an orange, a computer).
    b. **Object Detection:** Identifies and locates specific objects within an image or video.
    c. **Image Generation:** Creates new images or enhances existing ones.
    d. **Object tracking:** Follows or tracks the object once it is detected. This task can be used in autonomous vehicles, for instance to detect objects such as pedestrians, etc.
    e. **Content-based image retrieval:** Uses computer vision to search, browse and retrieve images from large data stores.
3) **Artificial Neural Networks:**
    a. **Deep Learning:** Utilizes deep neural networks with many layers to model complex patterns in data, often used in computer vision, NLP, and more.
    b. **Knowledge Representation and Reasoning:** Focuses on how to represent and manipulate knowledge to support logical reasoning and problem solving.
    c. **Fuzzy Logic:** Deals with uncertainty and imprecision, allowing AI systems to work with approximate or incomplete information.

Activities:

1. NLP Toolkit Workshop

Objective: To gain hands-on experience with NLP tools.

Activity: Students use NLP toolkits (like NLTK in Python) to perform tasks such as part-of speech tagging, sentiment analysis, and name entity recognition on sample texts.

2. AI Art Gallery

Objective: Explore the creative capabilities of AI in generating art.
Activity: Students use AI-based image generation tools to create artwork. They can then host a virtual art gallery, explaining the AI processes behind each piece.

3. Computer Vision Scavenger Hunt
   Objective: Understand the practical applications of computer vision.
   Activity: Students use their smartphones to capture images of various objects. They then use an AI-based image recognition app to identify these objects, noting the accuracy and any interesting findings.
4. AI Debate: Ethical Implications
   Objective: To discuss the ethical implications of AI capabilities.
   Activity: Organize a debate on topics such as privacy concerns in speech recognition, the impact of AI-generated art on human creativity, or the ethical use of facial recognition technology.
5. Interactive Neural Network Simulator
   Objective: To understand the workings of neural networks.
   Activity: Students use an online neural network simulator to build and train a basic model. They experiment with different architectures and observe the effects on performance.
6. Fuzzy Logic Case Studies
   Objective: To understand the application of fuzzy logic in AI.
   Activity: Students analyze case studies where fuzzy logic is applied, such as in consumer electronics or weather forecasting, and discuss its advantages over traditional binary logic.
7. AI in Movies: Fact vs. Fiction
   Objective: To differentiate between realistic and fictional portrayals of AI capabilities.
   Activity: Students watch clips from movies featuring AI and discuss which aspects are scientifically accurate and which are purely fictional, based on their knowledge of AI's capabilities.
8. AI-Assisted Creative Writing
   Objective: Explore the creative side of AI in language generation.
   Activity: Students use an AI-based text generator to write short stories or poems. They focus on how AI can assist in the creative process and the limitations of AI in understanding context and creativity.
9. AI Ethics Workshop: Bias in AI
   Objective: Understand and discuss the implications of bias in AI systems.
   Activity: Students participate in workshops where they learn about how biases can enter AI systems and the consequences thereof. They then brainstorm strategies to mitigate these biases.

Lesson 4: Synergistic Fields in Computational Intelligence

Evolutionary Algorithms

Evolutionary Algorithms are algorithms that are efficiently searching for solutions, inspired by Darwinian evolution. It solves problems by employing processes that mimic the behaviors of living things. These algorithms are strong and flexible, which makes them good at finding an optimal solution.

1) **Memetic computing**

   The term “meme” in memetic algorithms refers to a unit of information or behavior that can be passed from one individual to another. In the context of algorithms, memes are sets of instructions or strategies. These memes evolve over time through a process inspired by Darwin’s theory of evolution.

   Memetic algorithms are designed to explore and create new solutions. While they may start with random or existing solutions, the process involves combining and modifying these solutions to explore new possibilities. The idea is to find more effective strategies over time through a kind of “survival of the fittest” mechanism, where successful solutions are more likely to be passed on and modified.

2) **Genetic programming**

   Genetic programming is an evolutionary algorithm that can program themselves by imitating biological processes like evaluation and mutation. Genetic programming can solve complex problems that humans might not know how to solve directly. Through random mutation, crossover, a fitness function, and multiple generations, genetic programming evolves solutions, often outperforming human-derived solutions.

   Selection rules: Selects the parents from the current population will contribute to the creation of the next generation.

   Crossover rules: Combining genetic information from two parents to create children for the next generation.

   Mutation rules: Making random changes to individual parents to introduce genetic diversity in the population.

   Fitness function: Evaluates how well a given solution is to the optimum solution.

Activities:

1- Do the Jupyter notebook activities.

2- Use a simulation tool or software (e.g., DEAP library in Python) to perform Genetic Programming. https://deap.readthedocs.io/en/master/tutorials/advanced/gp.html

3- Explore the impact of hyperparameter tuning in Genetic Programming. Have students experiment with different selection, crossover, and mutation parameters.

4- Assign each student or group a specific domain (e.g., healthcare, finance, logistics). Research and present a case study where Genetic or Memetic Computing was applied to solve a problem in that domain.

Lesson 5: Summary of Safe, Secure, and Reliable AI

Artificial intelligence comes with a lot of benefits to it. However, it is important that AI is designed to be safe, secure, and reliable. Achieving this goal involves several steps, such as obtaining clean and unbiased data. The quality of the data used to train AI models has a big impact on their performance and ethical implications. Ensuring data safety enables AIs to be more accurate and be free from biases. Additionally making the code transparent is also important, so we can understand how AI makes decisions and the patterns it follows.

Machine learning and deep learning are also crucial when building safe, secure, and reliable AI. Machine learning models can be “poisoned” during the deployment process, and deep learning models also should be designed to be transparent, allowing developers to understand the patterns they follow.

Artificial intelligence is used in various branches and has many capabilities. such as Natural Language Processing (NLP), Computer Vision, and Artificial Neural Networks. NLP empowers AI to comprehend and generate human-like language, computer vision enables machines to interpret and interact with visual data, and artificial neural networks, including deep learning and fuzzy logic, contribute to modeling complex pattern and reasoning process.

Genetic and Memetic programming serve as nature’s problem solvers, adapting and changing over time to find solutions that might even outsmart what humans can come up with directly. Evolutionary algorithms are used for tuning parameters in machine learning models, helping to find the best settings for optimal performance.

In summary, building good AI involves using clean data, making sure the code is clear, and being cautious when using machine learning and deep learning. By doing these things, we create AI that is helpful, fair, and trustworthy, making our future with technology better for everyone.

References:

1) https://www-formal.stanford.edu/jmc/whatisai.pdf

2) https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

3) https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/3-sec-characteristics#:~:text=Safe%20operation%20of%20AI%20systems,deployers%20and%20end%20users%3B%20and

4) https://adversa.ai/blog/what-is-secure-trusted-ai/

5) https://scienceexchange.caltech.edu/topics/artificial-intelligence-research/trustworthy-ai#:~:text=In%20the%20case%20of%20AI,result%20in%20a%20stable%20landing.

6) https://cs.lbl.gov/what-we-do/machine-learning/secure-machine-learning/

7) https://www.nytimes.com/2022/04/05/technology/ai-voice-analysis-mental-health.html

8) https://www.nature.com/articles/s41598-020-73917-0

9) https://onlinedegrees.sandiego.edu/artificial-intelligence-education/#:~:text=The%20Potential%20Benefits%20of%20AI%20in%20Education,-Ideally%2C%20writes%20Lynch&text=AI%20systems%20easily%20adapt%20to,deliver%20customized%20support%20and%20instruction.%E2%80%9D

10) https://www.ibm.com/topics/natural-language-processing#:~:text=the%20next%20step-,What%20is%20natural%20language%20processing%3F,same%20way%20human%20beings%20can.

11) https://www.ibm.com/topics/computer-vision

12) https://www.sciencedirect.com/science/article/pii/S2210650211000691

13) https://geneticprogramming.com/

14) https://deap.readthedocs.io/en/master/tutorials/advanced/gp.html