Introduction

Background and Significance of AI in Decision Making

Introduction: Background and Significance of AI in Decision Making

This section provides a foundational understanding of AI's role in decision-making and highlights the potential pitfalls of over-reliance on AI systems.

Understanding the Rise of AI in Decision Making:

- **AI-powered decision support tools are increasingly prevalent:** AI is being integrated into various fields, from loan approvals to medical diagnoses, to assist humans in making informed decisions.
- **The promise of human+AI teams:** The initial expectation was that combining human and artificial intelligence would lead to superior performance compared to either working alone.
- **The problem of overreliance:** Studies reveal that people often over rely on AI suggestions, even when those suggestions are incorrect, leading to suboptimal outcomes.

The Limitations of Explainable AI (XAI):

- **XAI's intended purpose:** Explainable AI aims to provide insights into AI's reasoning, enabling users to identify situations where the AI's suggestions might be flawed.
- **XAI's shortcomings:** Despite the intention, XAI has not been entirely successful in reducing overreliance. People still tend to make poorer decisions when relying on incorrect or suboptimal AI suggestions.
- **The dual-process theory perspective:** Humans often rely on System 1 thinking (heuristics and shortcuts) rather than engaging in analytical thinking (System 2), making them vulnerable to cognitive biases.
- **The cognitive effort barrier:** Evaluating every AI explanation requires significant cognitive effort, which people tend to avoid.
- **Explanations as signals of competence:** Explanations can be misinterpreted as general indicators of AI competence, leading to increased trust and overreliance, regardless of the explanation's actual content.

Strategies for Mitigating Overreliance:

- **Focus on cognitive motivation:** To reduce overreliance, it's crucial to enhance people's cognitive motivation to engage analytically with AI explanations.
- **Employ cognitive forcing functions:** Cognitive forcing functions are interventions designed to disrupt heuristic reasoning and encourage analytical thinking. Examples include checklists, diagnostic time-outs, or explicitly ruling out alternatives.
- **Balance effectiveness and acceptability:** Stricter interventions might promote deeper thinking but could be perceived as complex and less usable.

Introduction: Background and Significance of AI in Decision Making

Understanding the Rise of AI:

- **Integration:** Artificial intelligence (AI) is increasingly present in everyday technology.
- **Opacity:** Algorithms on common platforms are often unclear to users. Many users don't realize they are interacting with AI.

- **Impact:** This lack of understanding limits the ability to effectively use, collaborate with, and critically evaluate AI.
- **Consequences:** Misconceptions about AI can lead to misdirected regulations and public disappointment if AI development expectations are not met.

Addressing Misunderstandings:

- **Black-box Algorithms:** Algorithms with obscured inner workings contribute to misunderstandings.
- **Lack of Technical Knowledge:** Even with transparent technologies, a lack of user knowledge can lead to misconceptions.
- **Need for Understanding:** There's a need to understand AI from both learner and designer perspectives.

Defining AI Literacy:

- **Competencies:** AI literacy encompasses the necessary skills for a future where AI transforms communication, work, and life.
- **Initiatives:** Companies and educators are broadening AI education to underrepresented audiences and incorporating AI into curricula.
- **AI for K12:** This working group is developing standards for K-12 classrooms, outlining what each grade band should know about AI.

Key Concepts of AI (The Five Big Ideas):

1. **Perception:** Computers perceive the world using sensors.
2. **Modeling and Reasoning:** Agents maintain models/representations of the world and use them for reasoning.
3. **Learning:** Computers can learn from data.
4. **Human-AI Interaction:** Making agents interact with humans is a substantial challenge for AI developers.
5. **Societal Impact:** AI applications can impact society in both positive and negative ways.

AI Literacy in Relation to Other Literacies:

- **Digital Literacy:** A prerequisite for AI literacy, as individuals need to understand how to use computers.
- **Computational Literacy:** Not necessarily a prerequisite, but understanding programming can aid in making sense of AI.
- **Scientific Literacy:** Can inform AI literacy, particularly understanding machine learning practices.
- **Data Literacy:** Closely related to machine learning, with overlapping competencies.

Defining AI Literacy:

- **Definition:** AI literacy is a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace.

Introduction: Background and Significance of AI in Decision Making

This section provides guidance on understanding and leveraging AI in your organization, focusing on creating a strategic approach to AI implementation.

1. Understand the Current Landscape:

- **Acknowledge the AI Evolution:** Recognize that AI technology is rapidly evolving, and many organizations are struggling to keep pace.
- **Recognize the Importance:** Understand that scaling AI or generative AI use cases to create business value is a top priority for many organizations.

2. **Develop an AI Playbook:**

- **Focus on Business Priorities:** Don't adopt AI just to follow trends. Clearly define the critical business problems you're trying to solve and how AI can contribute to your overall strategy.
  - **Deconstruct Challenges:** Break down business challenges into smaller subproblems to identify specific AI technologies and techniques that can be helpful.
  - **Separate Signal from Noise:** Abstract the essential information from irrelevant data.
- **Assess Data Readiness:** Ensure your organization's data is suitable for AI.
  - **Data Quality:** Launch data cleansing and data management initiatives.
  - **Data Governance:** Establish data governance policies.
  - **Third-Party Data:** Align with the right third-party data partners when necessary.
- **Evaluate Employee Skills:** Determine the AI maturity level of your employees and identify key skills gaps.
  - **Upskilling Programs:** Implement programs to upskill and train employees on technical skills and technical decision-making.
- **Foster a Supportive Culture:**
  - **Cross-Functional Teams:** Create silo-busting cross-functional teams.
  - **Embrace Failure:** Make failure permissible to encourage creativity.
  - **Human-Machine Collaboration:** Encourage innovative ways to combine human and machine capabilities.

3. **Manage Expectations:**

- **Adopt a Two-Tier Approach:** Implement a "fast-and-slow" strategy.
  - **Fast Experiments:** Conduct rapid experiments and proofs of concept.
  - **Slow Strategy:** Use the lessons and data from experiments to inform a slower, longer-term strategy.
- **Apply "Thinking Fast and Slow":** Use the concept of fast, intuitive thinking and slow, analytical thinking to enhance organizational capabilities and competitiveness.
- **Maintain a Practical Perspective:** Ensure enthusiasm for AI is grounded in a practical, reasonable strategy for integrating it into your business.

4. **Embrace Experimentation and Responsible AI:**

- **Continuous Experimentation:** Continuously experiment, build proofs of concept, and promote sandbox activities.
- **Capture Lessons:** Capture the lessons and data from both successes and failures to inform AI strategies.
- **Ensure Responsible Use:**
  - **Ethical Values:** Identify and communicate your ethical values.
  - **Accountability:** Create accountability through organizational structures.
  - **Bias Detection and Remediation:** Take continuous steps to detect and remediate bias.

How-To Guide: Introduction - Background and Significance of AI in Decision Making

This section provides background and highlights the significance of AI in modern decision-making processes.

1. Understanding the Shift: Data Takes Center Stage

- **Recognize the Increased Awareness:** Companies are now more aware of the power of data, not just as an enabler, but as a central asset for new business models, customer insights, governance, ethics, and risk management. Data has moved from being "just there" to being a critical focus.
- **Acknowledge Past Neglect:** Previously, data investment was often overlooked. Now, organizational agility depends on data's ability to facilitate rapid responses.
- **Embrace Agility:** Understand that agility comes from hyperawareness about data, informed decision-making, and fast execution. Invest in data management to improve performance.

2. The Duality of Data: Enabling Processes and Products

- **Data as an Enabler:** Data enables both processes and products/services. It has also become a product in its own right.
- **Consistent Approach:** Develop a consistent approach to managing data, considering its dual role.
- **Address Data Silos:** Break down data silos to create a comprehensive view of end-to-end value flow and generation. Extract data from various systems and consolidate it into a central repository (e.g., a data lake).

3. Bridging the Expectation Gap

- **Acknowledge Higher Expectations:** Recognize that expectations for data utilization are increasing rapidly.
- **Avoid Panic Actions:** Resist the urge to take drastic measures that could undermine previous data consolidation efforts.

4. Centralization vs. Decentralization: Finding the Right Balance

- **Consider End-to-End Processes:** Think in terms of end-to-end processes while also allowing for independent data usage.
- **Determine Centralization Needs:** Decide what data should be centralized and what should be decentralized to avoid creating new silos.
- **Optimize Data Lake Strategy:** Determine whether a single data lake or multiple data lakes are the best approach for efficient processes.

5. Adapting to the New World of Data

- **Overcome ERP Limitations:** Recognize that traditional ERP interfaces may not be user-friendly for working with data.
- **Democratize Data Access:** Bring more data to more employees to unlock its full value.
- **Embrace Real-Time Feedback:** Implement processes that allow for real-time feedback cycles to improve and transform data lifecycles and business processes. Avoid rigid, long-term processes.

6. Applying Physical Goods Management Principles to Data

- **Learn from Physical Goods Management:** Apply the disciplines learned in managing physical goods to the data value chain.
- **Focus on Simplicity for Agility:** Simplify and consolidate information to achieve agility.
- **Implement Data Governance:** Establish data governance to ensure data quality, consistency, and security.

Introduction: Background and Significance of AI in Decision Making

This section provides a historical overview of AI and its growing significance in decision-making, particularly within higher education.

Understanding the Roots of AI:

- **1956 Dartmouth Workshop:** This pivotal event marked the formal beginning of AI as a field. Key figures like John McCarthy, Arthur Samuel, Marvin Minsky, and Claude Shannon gathered to discuss the potential of machines to mimic human intelligence.
- **Early Challenges:** Initial limitations included computing power, data storage, and effective programming techniques.
- **Key Concepts Discussed:** The workshop addressed crucial concepts like automatic computers, computer language programming, neuron nets, calculation theory, self-improvement of data programs, abstractions, and randomness/creativity.

The Evolution of Computing and its Impact on Decision-Making:

- **Traditional Decision-Making (Pre-AI):** Decisions in universities were primarily based on the professional expertise and judgment of experienced individuals.
- **Administrative Computing Emerges (1950s-1960s):** Large mainframe computers were introduced, initially for academic purposes and later for administrative tasks.
- **Statistical Software (Late 1960s-1970s):** Tools like SPSS and SAS became available, enabling institutional researchers to analyze data and inform leadership.
- **Early AI Experiments in Education:**
  - **Project Essay Grade:** An early attempt to automate the grading of student essays using computer programs. While promising, it was initially too costly for widespread adoption.
- **Data Storage Advancements:**
  - **Punch Cards:** Early method for data storage.
  - **Magnetic Tape:** Introduced by IBM in 1951, offering increased speed and volume.
  - **Hard Disks:** Enabled random access to data, improving storage capacity and retrieval speed.
  - **Floppy Disks:** Facilitated data transport between microcomputers and mainframes.
- **Data Management Systems:**
  - **Navigational Databases (1960s):** Early forms of data organization.
  - **Relational Databases (1970s):** Developed by Edgar Codd, providing a more structured approach to data management.
  - **SQL (Structured Query Language):** Enabled efficient data retrieval and manipulation.
- **The Microcomputer Revolution (1980s):** Personal computers became affordable and widespread, empowering individuals to generate and access data.

Introduction: Background and Significance of AI in Decision Making

Understanding the Role of AI in Modern Education

Artificial Intelligence (AI) is rapidly changing the landscape of higher education. Generative AI tools, such as ChatGPT, are becoming increasingly important for:

- **Enhancing Curriculum Development:** AI can assist in creating personalized and adaptive educational content.
- **Boosting Student Engagement:** AI tools can make learning more interactive and tailored to individual needs.

Key Benefits of AI Integration

- **Personalized Learning:** AI technologies enable personalized learning experiences through chatbots, predictive analytics, and automation.
- **Data-Driven Adjustments:** AI provides insights into student performance, allowing for data-driven course adjustments.
- **Improved Efficiency:** AI can streamline curriculum development, leading to quicker content production.
- **Inclusive Content:** AI facilitates the creation of course materials suited to varied learning styles, enhancing flexibility and inclusivity.

How AI Enhances Teaching and Learning

- **Administrative and Teaching Tool:** AI serves as both an administrative aid and a teaching tool, enhancing scalability and cost-effectiveness.
- **Personalized Assistance:** AI enhances teaching by providing automated support and tailoring curricula to individual student needs.
- **Skill Development:** AI promotes skill development in communication and collaboration.
- **Improved Learning Environments:** AI integration leverages data to create improved learning environments and outcomes, fostering student engagement.

Practical Applications of Generative AI

- **Personalized Learning Pathways:** Generative AI revolutionizes education through personalized learning pathways, increasing student engagement and success.
- **Continuous Adaptation:** AI enables faculty to continuously adapt and enhance course materials.
- **Innovative Course Development:** AI facilitates innovative course development by integrating tools like speech recognition, plagiarism detection, and video editing.
- **Virtual Tutors:** AI transforms the vision of a virtual tutor, accessible anytime, into reality through chatbot systems.
- **Adaptive Assessments:** AI personalizes student assessments and crafts questions that adapt to unique learning styles, increasing the accuracy and inclusivity of evaluations.
- **Timely Feedback:** AI offers immediate, tailored feedback crucial for mastery learning.

Balancing AI with Traditional Methods

While AI offers numerous benefits, it's important to balance its use with traditional teaching methods and address challenges such as academic integrity and intellectual development.

Introduction: Background and Significance of AI in Decision Making

This section provides a guide to understanding and addressing overreliance on AI in decision-making, particularly focusing on how explanations can be used effectively.

Understanding Overreliance:

- **Definition:** Overreliance occurs when people accept an AI's decision, even when it's incorrect. This is a common problem in human-AI collaboration.
- **Why it Matters:** Overreliance can lead to poor decisions, reinforce machine bias, and diminish human accountability.
- **The Role of Explanations (XAI):** The goal of Explainable AI (XAI) is to provide explanations for AI predictions, with the expectation that these explanations will help people identify and correct AI errors.

Why Explanations Sometimes Fail:

- **Cognitive Costs:** Both tasks and explanations have cognitive costs (difficulty). If the task is easy, people may rely on the AI without thinking. If the explanation is as difficult to understand as doing the task manually, people may skip it.
- **Trust and Anchoring:** The mere presence of an explanation can increase trust in the AI, even if the explanation is flawed. Explanations can also "anchor" people to the AI's prediction, making them less likely to question it.

How to Make Explanations Effective:

- **Cost-Benefit Analysis:** People strategically decide whether to engage with an AI explanation based on a cost-benefit analysis. They weigh the effort required to understand the explanation against the effort of completing the task alone and the risk of relying on a potentially incorrect AI.
- **Reduce Cognitive Costs:**
    - **Task Difficulty:** When tasks are more difficult, people are more likely to engage with explanations.
    - **Explanation Difficulty:** Make explanations easy to understand. The easier it is to verify the AI's reasoning, the more likely people are to use the explanation and correct errors.
- **Increase Benefits:** Incentives, such as monetary compensation, can encourage people to engage with explanations.

Key Takeaway:

Explanations can reduce overreliance, but only when they sufficiently reduce the cost of verifying the AI's prediction compared to the cost of completing the task alone. Focus on making explanations easy to understand and ensuring that the task itself is sufficiently challenging to motivate engagement.

Introduction: Background and Significance of AI in Decision Making

This section provides a foundational understanding of AI's role in modern decision-making processes.

Why AI Matters in Decision Making:

- **Speed and Efficiency:** AI significantly accelerates the analysis of large datasets, providing rapid insights for quicker decision-making, especially under tight deadlines.
- **Improved Accuracy:** AI can enhance the accuracy of data analysis and reporting, leading to better business predictions and strategic planning.
- **Business Transformation:** AI facilitates the transformation of business strategies, enabling quick adjustments and improved productivity.
- **Cost Reduction:** AI optimizes operations, predicts demands, and adjusts supply levels, leading to noticeable reductions in operating costs.
- **Data-Driven Choices:** AI software empowers companies to make data-driven decisions, improve customer targeting, and enhance development.
- **Customization:** AI assists in customizing data and reports for speedy and precise strategic decisions.

How AI Impacts Internal Operations:

- **Departmental Applications:** AI transforms operations across various departments, including transport, consultation management, Human Resources, Accounts & Finance, Sales and Marketing.
- **Demand Prediction:** AI predicts demands and fluctuations, enabling proactive adjustments to supply and data levels.
- **Prioritization and Optimization:** AI prioritizes and optimizes results, streamlining processes and improving efficiency.

Key AI Mechanisms:

- **Automation Efficiency:** AI automates tasks, freeing up human resources for more strategic activities.
- **Complex Management:** AI handles multilayered problems that would overwhelm human policymakers.
- **Succession Planning Incorporation:** AI can assist in identifying and developing future leaders.
- **Risk Monitoring and Assessment:** AI aids in identifying and assessing potential risks.
- **Compliance and Scam Recognition:** AI helps ensure compliance with regulations and detects fraudulent activities.

Industry Applications:

- AI is applicable across various industries, including Healthcare, Finance, Consulting Business, Transportation, and Food & Beverage.
- Government authorities can leverage AI in defining and implementing automation processes.

Important Considerations:

- **Ethical Concerns:** Address ethical considerations related to AI implementation.
- **Algorithm Bias:** Ensure nonbiased and fair decision-making algorithms.
- **Transparency and Accountability:** Prioritize transparency and accountability in AI-driven processes.
- **Data Privacy Protection:** Implement robust data privacy protection measures.
- **Explainability:** Address explainability issues in AI decision-making.
- **Human-AI Collaboration:** Recognize the importance of Human-AI association.

Addressing Challenges:

- **Data Quality:** Ensure the quality of input data to avoid adverse effects on reports and decisions.
- **Data Clarity:** Ensure data clarity for proper explanation of results.
- **Black Box Issue:** Address the "Black box issue" in AI by using typical-skeptical techniques to justify results.

Introduction: Background and Significance of AI in Decision Making

This section provides an overview of the increasing role of Artificial Intelligence (AI) in decision-making processes, particularly within the context of higher education.

Understanding the Rise of AI:

- **AI-Powered Tools:** Become familiar with AI tools based on advanced language models, such as ChatGPT and similar platforms (e.g., Bard). These tools are rapidly changing various sectors, including education.

- **Generative AI:** Understand the concept of generative AI and its implications for teaching and learning. Generative AI refers to AI models that can create new content, such as text, images, and audio.
- **ChatGPT as a Case Study:** Use ChatGPT as a specific example to explore the broader impact of AI on education. Analyze the conversations and debates surrounding its use in academic settings.

Exploring the Impact of AI in Education:

- **Revolutionizing Higher Education:** Recognize that AI has the potential to revolutionize higher education globally.
- **Innovative Pedagogies:** Consider how AI can be used to develop new and innovative teaching methods.
- **Curriculum Development:** Explore the potential of AI in decolonizing the curriculum and making it more inclusive.
- **Ethical, Legal, and Social Implications:** Address the ethical, legal, and social issues related to the use of generative AI in education.

Practical Strategies for Educators:

- **Stay Informed:** Keep up-to-date with the latest developments in AI and its applications in education.
- **Investigate AI Tools:** Explore different AI tools and platforms to understand their capabilities and limitations.
- **Engage in Discussions:** Participate in conversations and debates about the ethical and practical considerations of using AI in the classroom.
- **Seek Guidance:** Consult with experts and colleagues to learn about best practices for integrating AI into teaching and learning.
- **Harness the Power of AI Responsibly:** Use AI tools in a way that is ethical, effective, and promotes inclusivity.

Introduction: Background and Significance of AI in Decision Making

This section provides a foundational understanding of AI's role in decision-making, particularly within higher education.

Understanding AI Fundamentals:

- **Definition:** AI simulates human intelligence using computers and machines.
- **Core Components:** AI relies on algorithms and software to process data, learn from it, and make decisions or take actions.
- **Key Techniques:** AI encompasses various techniques, including:
  - Machine Learning
  - Deep Learning
  - Natural Language Processing
  - Computer Vision

AI's Impact Across Industries:

- **Healthcare:** AI assists in disease diagnosis and predicting patient outcomes.
- **Business & Finance:** AI is used for fraud detection, trading strategies, and personalized financial recommendations.
- **Transportation:** AI powers self-driving vehicles, traffic management systems, and predictive maintenance.

- **Retail:** AI provides customer insights, demand forecasting, and supply chain optimization.
- **Virtual Assistants:** AI enables chatbots and image generators through natural language processing and computer vision.

AI in Higher Education:

- **Transformative Potential:** AI is changing how students learn and instructors teach.
- **Applications:**
  - Personalized learning experiences
  - Advanced data analysis for student success (e.g., identifying struggling students)
  - Enhanced retention rates
  - Improved overall quality of education

Historical Integration of AI in Higher Education:

- **Early 2000s:** Integration into Learning Management Systems (LMSs) for course organization and assessment.
- **2010s:** Personalized learning through AI algorithms recommending courses and content. Adaptive learning systems adjusted content difficulty based on student progress.
- **Mid-2010s:** Chatbots and virtual assistants provided 24/7 support for student inquiries and administrative processes.
- **Late 2010s:** Predictive analytics identified at-risk students, enabling targeted support.
- **2020s:** AI-powered proctoring tools for online exams, AI-generated content, and automated grading systems.
- **Present:** Exploration of AI to enhance the overall campus experience, including AI-driven lifecycle management systems for students.

Key Considerations:

- **Data Privacy:** Be aware of the implications of using student data.
- **Biases:** Recognize and address potential biases in AI algorithms.
- **Ethical Considerations:** Engage in ongoing discourse about the ethical implications of AI in education.
- **Faculty and Student Readiness:** Assess and address the readiness of faculty and students for AI integration.

Thesis Statement and Research Questions

Thesis Statement and Research Questions: A Guide

1. Identifying Key Areas of Impact:

- **Focus:** Consider the broad societal implications of AI. Think about how AI affects:
  - Labor and Automation
  - Bias and Inclusion
  - Rights and Liberties
  - Ethics and Governance

2. Formulating Research Questions:

- **Labor and Automation:**
  - How is AI used in employee hiring, firing, and management, including surveillance?
  - Are robots and automated systems replacing or complementing human workers?
  - How does AI transform the nature of work itself?
- **Bias and Inclusion:**
  - How do biases manifest in AI decision-making systems?

    - What are the social implications of biased AI?
- **Rights and Liberties:**
    - How do AI systems support or erode citizens' rights and liberties in areas like criminal justice, law enforcement, housing, hiring, and lending?
    - How can AI enhance transparency and accountability in law enforcement (e.g., with police body cameras)?
- **Ethics and Governance:**
    - Whose concerns are reflected in the ethics of AI?
    - How can ethical codes and strategies be developed for AI?
    - How can we ensure public discourse and the ability to opt-out of AI systems?

3. Understanding Terminology:

- **Bias:** Be aware of the different meanings of "bias" in popular usage, legal contexts, and machine learning/statistics.
    - **Popular:** Judgement based on preconceived notions or prejudices.
    - **Legal:** Lack of impartiality.
    - **ML/Statistics:** Difference between an estimator's expected value and the true value; selection bias.
- **Algorithms, Data, & AI (ADA):** Understand the core components:
    - **Algorithm:** An unambiguous procedure for solving a problem.
    - **Data:** Encoded information about target phenomena.
    - **AI:** Technology that performs tasks considered intelligent.

4. Addressing Tensions Between Values:

- **Recognize Conflicts:** Be aware of potential conflicts between values when developing and deploying AI. Examples include:
    - Quality of services vs. privacy
    - Personalization vs. solidarity
    - Privacy vs. transparency
    - Convenience vs. dignity
    - Accuracy vs. explainability
    - Accuracy vs. fairness
    - Preferences vs. equality
    - Efficiency vs. safety

Introduction: Thesis Statement and Research Questions

While this document doesn't explicitly provide instructions on *how* to write a thesis statement or research questions, it offers insights into *what* those elements should achieve within the context of a research project focused on law, ethics, and technology. Consider these points when formulating your thesis and research questions:

- **Focus and Scope:** Define a clear and manageable scope. Acknowledge the limitations of your research, especially when dealing with broad topics or multiple jurisdictions.

- **Identify Key Issues:** Pinpoint the central problems or areas of interest. Examples from the document include:

    - Algorithmic fairness and explainability.
    - Autonomous vehicles.
    - Effects of AI and robotics on the labor market.
    - Ownership of works created by AI.
    - Legal status of robots (e.g., personhood).

    - Safety and civil liability issues related to robotics.
- **Highlight Gaps and Challenges:** Identify areas where existing laws, regulations, or understanding are lacking. This can form the basis of your research questions.
- **Consider the Role of Non-Governmental Actors:** Acknowledge the influence of private industry standards and guidelines in shaping the development and regulation of AI and robotics.
- **Address Convergences and Distinctions:** If conducting comparative research, aim to identify both the commonalities and differences in how AI and robotics are regulated across different jurisdictions.
- **Reflect Current Developments:** Ensure your research questions and thesis are relevant to the current state of the field, acknowledging recent legal developments, case law, and proposed legislation.

Thesis Statement and Research Questions

Crafting a strong thesis statement and relevant research questions is crucial for any ethical inquiry. Here's how to approach it:

1. Identifying Ethical Dilemmas:

- **Explore Diverse Fields:** Ethical questions aren't confined to philosophy. Consider ethical implications within various disciplines like business, biology, communication, education, engineering, nursing, political science, psychology, sociology, social work, and even mathematics.
- **Brainstorm:** Prompt your thinking by considering real-world scenarios and potential conflicts of values or principles.
- **Adapt Existing Questions:** Review sample ethical questions from different fields and adapt them to your specific area of interest. A slight tweak can make a question relevant to a new context.

2. Formulating Research Questions:

- **Focus on Specific Issues:** Frame your questions around specific ethical issues within your chosen discipline.
- **Consider Stakeholders:** Think about how different individuals or groups (stakeholders) might be affected by the ethical dilemma.
- **Explore Underlying Factors:** Investigate the factors that contribute to or influence the ethical issue.
- **Aim for Open-Ended Questions:** Avoid questions with simple "yes" or "no" answers. Instead, formulate questions that encourage exploration and analysis.

3. Developing a Thesis Statement:

- **Summarize Your Position:** Your thesis statement should clearly state your position or argument regarding the ethical issue.
- **Provide a Roadmap:** It should provide a roadmap for your research, outlining the key points you will explore.
- **Ensure it is Debatable:** A strong thesis statement presents a viewpoint that others could reasonably disagree with.
- **Connect to Research Questions:** Ensure your thesis statement directly addresses the research questions you have formulated.

Examples of Ethical Questions by Discipline:

- **Business/Accounting/Marketing:** How can regulators determine the ethical nature of a financial transaction?
- **Biology:** What are the ground rules for conducting and reporting biological research in an ethical manner?
- **Communication Studies:** How does the pressure to create a steady stream of "clickable content" change the ways journalists conceptualize, gather, report and produce news?
- **Education:** Does assistive technology provide ethically defensible educational support to a student with mild behavioral problems, or does it represent an unfair advantage?
- **Engineering:** How much exposure do engineering students receive to ethics in their courses of study?
- **Nursing:** What influences are uppermost in the ethical decision-making processes of student nurses as they engage with patients?
- **Political Science:** What ethical responsibility do educators bear for the actions of their students?
- **Psychology/Sociology:** In what ways do sports create and/or reflect ethical dilemmas that relate to broader psychological and sociological issues?
- **Social Work:** How do social workers and social work educators in the U.S. interpret and mobilize the section of the NASW Code of Ethics addressing social workers' ethical commitments to broader society?
- **Mathematics:** How do one's personal axioms shape one's world both mathematically and personally, if an axiom is taken to be a statement believed without proof?

Thesis Statement and Research Questions

This document doesn't explicitly provide instructions on writing a thesis statement or research questions. However, it presents a scenario that can be used as a basis for formulating them. Here's how you can approach it:

1. **Identify the Core Problem:** Understand the central issue at the heart of the case study. In this case, it's the city of New Leviathan's struggle with rising crime, racial tensions, and budgetary deficits, and the proposed use of AI-driven data analytics to address these problems.

2. **Formulate a Thesis Statement:** A thesis statement should present a clear argument or position regarding the core problem. Consider the ethical implications of using AI in public sector data analytics.

   - **Example Thesis:** "While AI-driven data analytics offers a potential solution to New Leviathan's crime problems, its implementation raises significant ethical concerns regarding privacy, bias, and accountability that must be carefully addressed to ensure equitable and just outcomes."
3. **Develop Research Questions:** Research questions should explore specific aspects of the problem and guide your investigation.

   - **Example Research Questions:**
     - "What are the potential biases embedded in the data used to train the crime forecasting algorithms, and how might these biases disproportionately affect certain communities within New Leviathan?"
     - "What measures can be implemented to ensure transparency and accountability in the use of AI-driven data analytics for crime reduction in New Leviathan?"

        - "How can the city of New Leviathan balance the potential benefits of AI-driven crime reduction with the need to protect the privacy rights of its citizens?"
        - "What are the potential unintended consequences of using social network analysis to predict criminal behavior, and how can these consequences be mitigated?"

4. **Consider Different Perspectives:** The case study presents multiple stakeholders (Mayor Hobbes, WCG, citizens of New Leviathan, NLPD). Consider their perspectives when formulating your thesis and research questions.

5. **Focus on Ethical Implications:** Given the context of AI ethics, your thesis and research questions should heavily emphasize the ethical considerations surrounding the use of AI in the public sector.

Thesis Statement and Research Questions: A Guide

This section focuses on how to formulate a strong thesis statement and relevant research questions when exploring the intersection of human cognition and AI-assisted decision-making.

Understanding the Problem: Overreliance on AI

- **Identify the Issue:** Recognize that people often overrely on AI suggestions, even when those suggestions are incorrect. This can lead to poorer decisions compared to situations where they would have made decisions without AI assistance.
- **Question the Effectiveness of Explanations:** Be aware that simply providing explanations for AI recommendations (Explainable AI or XAI) may not solve the overreliance problem. In some cases, explanations might even increase trust and overreliance without improving decision-making.

Leveraging Dual-Process Theory

- **Acknowledge Cognitive Processes:** Understand that humans primarily use "System 1" thinking, which relies on heuristics and mental shortcuts. While efficient, this can lead to cognitive biases and suboptimal decisions.
- **Recognize the Limitation of Analytical Engagement:** Realize that people are unlikely to engage analytically with every AI explanation due to the cognitive effort required. They may instead form general impressions of the AI's competence.

Formulating Your Thesis

- **Address the Need for Cognitive Motivation:** Argue that reducing overreliance requires not only effective explanation techniques but also methods to increase people's motivation to analyze explanations critically.
- **Consider Cognitive Forcing Functions:** Explore cognitive forcing functions as potential interventions. These functions disrupt heuristic reasoning and encourage analytical thinking at the point of decision-making. Examples include checklists, diagnostic time-outs, or explicitly ruling out alternatives.
- **Acknowledge Usability and Acceptability:** Recognize the importance of usability and acceptability when designing cognitive forcing functions. Stricter interventions might be more effective but could be perceived as complex and less usable.

Developing Research Questions

- **Investigate the Effectiveness of Cognitive Forcing:** Frame questions around whether cognitive forcing functions can reduce overreliance on AI compared to simple explainable AI approaches.
- **Explore the Trade-off Between Effectiveness and Acceptability:** Ask questions about the relationship between the effectiveness of interventions and their perceived difficulty, preference, and trustworthiness.
- **Consider Individual Differences:** Investigate how individual differences, such as "Need for Cognition" (NFC), moderate the effectiveness of interventions. Do people with high NFC benefit more from cognitive forcing functions?
- **Address Potential Inequalities:** Frame questions around whether interventions benefit all users equally, or if they disproportionately benefit certain groups (e.g., those with high NFC).

Introduction: Thesis Statement and Research Questions

This guide section focuses on developing your research and presenting it effectively in poster format. The assignment is divided into two deliverables, each requiring a poster presentation:

Deliverable 1: "I, an AI Researcher"

This deliverable focuses on exploring AI research labs and industries.

1. **Choose a Topic:** Select one of the following areas:
    - AI Research Laboratories (outside academia)
    - Industries that use AI
    - AI Research Laboratories at R1/R2 Universities (Ph.D. positions)
2. **Identify Examples:** Find 5-10 examples related to your chosen topic (these can be international).
3. **Justify Your Choices:** Explain why you selected your chosen topic and specific examples.
4. **Investigate Your Examples:** For each example, research the following:
    - What do they do?
    - What is their mission?
    - How big are their research teams?
    - What is their location?
    - What kinds of AI problems do they focus on?
    - What impacts do they seem to cause?
    - Are they multidisciplinary?
    - How diverse are their teams?
    - What are the requirements to be part of their team?
    - Would you like to be on their team? Why?
5. **Bridge Your Examples:** Analyze the commonalities and differences between your examples.
    - What do they have in common?
    - How do they differ?
    - Can you identify any trends?
    - What research directions seem to cause more impact?
    - How did this activity change the way you see AI research?
    - Translate your findings into diagrams.

6. **Create a Poster:**

   - Focus on the analysis from step 5.
   - Include at least one diagram.
   - Use LaTeX if possible for poster creation.

Deliverable 2: "The Same AI, Distinct Impacts" (GPT-3)

This deliverable explores the potential impacts of GPT-3 in a developing country.

1. **Understand GPT-3:** Research what GPT-3 is and its capabilities.

2. **Analyze Potential Impacts:** Understand how GPT-3 can advance both harmful and beneficial applications of language models.

3. **Choose a Developing Country:** Select a developing country.

4. **Investigate the Chosen Country:** Research the country's culture, politics, environment, and economics.

5. **Analyze Impacts in the Chosen Country:**

   - **Person 1:** Investigate and reflect on how GPT-3 can be used in harmful ways in the chosen country.
   - **Person 2:** Investigate and reflect on how GPT-3 can be used in beneficial ways in the chosen country.
   - **Person 3:** Support Persons 1 and 2, provide insights, and bridge the findings. Focus on poster design and ensure a cohesive presentation.

6. **Create a Poster:**

   - Focus on the analysis of GPT-3's potential impacts in the chosen country.
   - Include at least one diagram.
   - Use LaTeX if possible for poster creation.
   - Add a "Final Remarks" section summarizing what you learned from both deliverables.

General Poster Guidelines (Both Deliverables):

- **Digital Format:** Do NOT print your poster. Only the digital version will be used.
- **Presentation:** You will have 8 minutes to present your poster and findings to the class. Deliverable 2 includes 2 minutes for questions.
- **Visual Communication:** An effective poster is a visual communication tool.
- **Diagrams:** Include at least one diagram in each poster.
- **LaTeX:** LaTeX is highly preferable for creating posters.

Resources:

- Make a Good Poster
- An effective poster is a visual communications tool.
- The untold story of GPT -3 is the transformation of OpenAI
- Vivero Digital Project Peer Mentors . The Digital Liberal Arts Collaborative, the Libraries, and the Center for the Humanities are offering virtual drop- in hours in Teams, to support you with digital liberal arts assignments. “Peer Mentors can help you plan the early stages of a project, resolve issues along the way, or give feedback on design”. See times and additional details at http://bit.ly/ViveroDropInSpring2021 .
- Language Models are Few -Shot Learners

- GPT-3: Language Models are Few -Shot Learners
- How to Make Artificial Intelligence More Democratic
- GPT-3
- Understanding the AI alignment problem
- What does it take to create a GPT -3 product?
- Analyzing Machine -Learned Representations: A Natural Language Case Study

Thesis Statement and Research Questions

This document does not explicitly provide instructions on writing a thesis statement or research questions. Therefore, no practical information can be extracted for this subtopic.

Introduction: Thesis Statement and Research Questions

This section focuses on crafting a strong thesis statement and relevant research questions when exploring human trust in Artificial Intelligence (AI).

1. Defining the Scope:

- **Focus on Weak or Narrow AI:** Acknowledge that current AI is "weak" or "narrow," meaning it's designed for specific tasks, not general human-level intelligence. Frame your research questions and thesis statement accordingly.

2. Identifying Key Elements:

- **AI Representation:** Consider the physical form of the AI (robot, virtual agent, embedded system).
- **Machine Intelligence Level:** Acknowledge the varying capabilities of AI systems. Focus on *perceived* machine intelligence from the user's perspective.
- **Trust Dimensions:** Differentiate between cognitive and emotional trust. Cognitive trust relates to AI's reliability and competence, while emotional trust involves feelings of connection and security.

3. Formulating Research Questions:

- **AI Representation and Trust:** How does the physical representation of AI (robot, virtual, embedded) influence the development of cognitive and emotional trust?
- **Machine Intelligence and Trust:** How does the perceived level of machine intelligence affect human trust in AI?
- **Cognitive Trust Factors:** What specific factors (tangibility, transparency, reliability, task characteristics, immediacy behaviors) contribute to cognitive trust in AI?
- **Emotional Trust Factors:** What specific factors (tangibility, anthropomorphism, immediacy behaviors) contribute to emotional trust in AI?

4. Developing a Thesis Statement:

- **Integrate Key Elements:** Your thesis should connect AI representation, machine intelligence, and the development of cognitive and/or emotional trust.
- **Example:** "Human trust in AI is significantly influenced by the AI's physical representation and perceived level of machine intelligence, with tangibility, transparency, and reliability being key drivers of cognitive trust, and anthropomorphism playing a crucial role in emotional trust."

5. Considering the Broader Context:

- **Organizational Impact:** Acknowledge the potential impact of AI on workforce structure, job design, decision-making, and knowledge management.

- **Human-AI Collaboration:** Frame your research within the context of facilitating effective human-AI collaboration.

Thesis Statement and Research Questions: A Guide

1. Crafting Your Thesis Statement:

- **Focus:** Your thesis statement should clearly articulate the central argument or point you intend to make in your work.
- **Specificity:** Ensure your thesis is specific and focused.
- **Relevance:** The thesis should directly address the core topic of your research.

2. Developing Research Questions:

- **Purpose:** Research questions guide your investigation and help you explore specific aspects of your topic.
- **Alignment:** Ensure your research questions directly relate to and support your thesis statement. They should help you gather evidence and insights to support your central argument.
- **Clarity:** Formulate clear and concise research questions.
- **Number:** Aim for a manageable number of research questions.
- **Exploration:** Use research questions to delve into different facets of your topic.
- **Adaptability:** Be prepared to refine your research questions as your understanding of the topic evolves.

3. Connecting Thesis and Questions:

- **Relationship:** Your research questions should logically stem from your thesis statement. They break down the broader argument into smaller, investigable components.
- **Evidence:** Consider how each research question will contribute to gathering evidence that supports your thesis.

4. Refining Your Approach:

- **Revisit:** Regularly revisit and refine your thesis statement and research questions as you progress through your research. This iterative process ensures that your work remains focused and relevant.
- **Consider:** Think about the scope of your research and adjust your thesis and questions accordingly.
- **Ensure:** Ensure your research questions are answerable through investigation and analysis.

5. Seeking Feedback:

- **Share:** Share your thesis statement and research questions with others to get feedback on their clarity, focus, and relevance.

Thesis Statement and Research Questions

When developing a research project involving AI, particularly in the context of human subjects, clearly defining your thesis statement and research questions is crucial for IRB approval. Here's how:

1. Defining Research:

- **Determine if your project constitutes research:** Research is defined as a systematic investigation (including development, testing, and/or evaluation) designed to contribute to generalizable knowledge.

- **Systematic Investigation:** Ensure your project involves a careful or detailed inquiry or examination of information that involves a system, method, or plan.
- **Generalizable Knowledge:** Determine if the intended use of the AI is limited to the original dataset (non-generalizable) or intended as a tool to be applied to a broad population or not-yet-collected data (generalizable). If it is generalizable, it is more likely to be considered research.

2. **Human Subjects Involvement:**

- **Determine if your research involves human subjects:** A human subject is a living individual about whom an investigator is conducting research and either:
    - Obtains information or biospecimens through intervention or interaction with the individual, and uses, studies, or analyzes the information or biospecimens; or
    - Obtains, uses, studies, analyzes, or generates identifiable private information or identifiable biospecimens.
- **AI Tool Development:** If your research involves developing AI tools using data from human subjects, it likely requires IRB review.

3. **Formulating Research Questions:**

- **Clinical Investigations vs. Clinical Evaluations:**
    - **Clinical Investigations (Clinical Trials):** Frame your research questions around what works and doesn't work in treating humans. Focus on establishing safety, device performance, benefits, and effectiveness.
    - **Clinical Evaluations (Non-interventional assessments of existing data):** Frame your research question around whether the medical device can achieve its purpose. Focus on establishing safety and whether benefits outweigh risks.
- **Consider the Impacted Populations:** Ensure your research questions are relevant to the populations impacted by the findings.
- **Data Considerations:** Your research questions should consider the data set's size and diversity for the target population.

4. **IRB Considerations:**

- **Equitable Selection of Subjects:** Avoid unnecessary inclusion/exclusion of certain groups due to inconvenience.
- **Informed Consent:** Consider whether consent was obtained for future use of data in the manner proposed by your research questions. Determine if a waiver of consent is appropriate.
- **Data Monitoring:** Your research questions should address model iterations, data shift, version changes, and post-deployment monitoring to identify harms and mitigate biases.
- **Confidentiality:** Address how datasets will be maintained confidentially and how re-identifiability is minimized.
- **Privacy:** Identify what PHI/PII will be used and by whom, and consider additional protections for small populations and sensitive data.

Literature Review

Theoretical Frameworks of Human Decision Making

Theoretical Frameworks of Human Decision Making

Contemporary research on clinical reasoning emphasizes a dual-process model, comprising an intuitive, rapid Type 1 process and a deliberate, analytical Type 2 process (Norman et al., XXXX). Type 1 thinking is characterized as automatic, unconscious, and effortless, relying on heuristics and pattern recognition, while Type 2 thinking is controlled, conscious, and effortful, involving systematic analysis and logical reasoning (Evans & Stanovich, 2013).

Early studies in clinical reasoning, dating back to the 1970s and 1980s, revealed that clinicians often generate diagnostic hypotheses early in patient encounters, subsequently testing these hypotheses through data gathering (Gruppen et al., 5; Pelaccia et al., 6). This aligns with the dual-process model, where initial hypotheses are formed intuitively (Type 1) and then evaluated analytically (Type 2).

A key debate within this framework concerns the origins of errors in decision-making. One perspective attributes errors primarily to cognitive biases arising from Type 1 processing, suggesting that failures occur when Type 2 reasoning fails to detect and correct these biases (Kahneman, 2011). Conversely, another view posits that errors can originate from both Type 1 and Type 2 processes (Evans & Stanovich, 2013), with Type 2 errors potentially stemming from limitations in working memory capacity, which constrains computational processes.

The architecture of memory plays a crucial role in dual processing. Type 1 processing involves direct associations between new information and similar examples stored in long-term, associative memory (Ericsson & Pool, 2016). The likelihood of retrieving a similar example depends on the strength of the association, influenced by factors such as frequency, number of examples, and common features (Norman et al., XXXX). Type 2 processing, on the other hand, relies on computations within working memory, such as identifying features and estimating probabilities, placing a significant burden on limited cognitive resources (Evans & Stanovich, 2013).

Theoretical Frameworks of Human Decision Making: A Bayesian Perspective

This section explores the application of Bayesian statistical inference as a theoretical framework for understanding human decision-making, particularly in the context of cognitive models. Bayesian methods offer a principled approach to relating psychological models to empirical data, addressing limitations inherent in traditional frequentist approaches.

Key Arguments and Findings:

- **Bayesian Inference as a Comprehensive Approach:** The paper advocates for Bayesian inference as a complete and coherent method for connecting models and data in psychology. This approach contrasts with traditional frequentist methods, which have recognized limitations in parameter estimation and model selection.
- **Beyond Bayes Factors:** The paper argues against a narrow focus on Bayes Factors as the sole measure of model evaluation in Bayesian analysis. It emphasizes the importance of a full Bayesian analysis to address theoretical and empirical questions.
- **Model Complexity and Bayesian Inference:** A significant advantage of Bayesian inference is its ability to handle model complexity in a principled manner. The framework inherently accounts for the functional form effects of model complexity, particularly when varying parameters that influence parametric interactions.
- **Probabilistic Graphical Models:** The paper utilizes probabilistic graphical models to represent cognitive models. These models use nodes to represent variables of interest and

graph structures to indicate dependencies between variables. This approach facilitates the visualization and analysis of complex relationships within the models.

- **Application to Signal Detection Theory (SDT):** The paper applies Bayesian methods to a Signal Detection Theory (SDT) model of decision-making. This application demonstrates the ability of Bayesian methods to provide useful answers to important theoretical and empirical questions within the domain of decision-making.

Methodology:

- **Recasting Models as Probabilistic Graphical Models:** The study recasts existing cognitive models, such as the Signal Detection Theory model, into probabilistic graphical models. This involves defining variables of interest as nodes and specifying dependencies between them using the graph structure.
- **Posterior Sampling:** The Bayesian models rely on posterior sampling from graphical models. This involves estimating the posterior probability distribution of model parameters given the observed data.
- **Prior Specification:** The Bayesian approach requires the specification of prior distributions for model parameters. The choice of prior distributions can influence the posterior distribution and, consequently, the inferences drawn from the model.

Conclusions:

The paper concludes that Bayesian inference offers a powerful and versatile framework for analyzing cognitive models and understanding human decision-making. Its ability to handle model complexity, incorporate prior knowledge, and provide probabilistic inferences makes it a valuable tool for advancing psychological theory and empirical research. The application of Bayesian methods to models like Signal Detection Theory demonstrates its potential for addressing important theoretical and empirical questions in the field of decision-making.

Theoretical Frameworks of Human Decision Making

This research investigates the phenomenon of overreliance on AI in AI-assisted decision-making, where individuals accept incorrect AI suggestions even when they possess the knowledge to make a better choice independently. The study leverages the dual-process theory of cognition to explain this overreliance. This theory posits that humans primarily operate using System 1 thinking, which relies on heuristics and cognitive shortcuts. While efficient for routine decisions, System 1 thinking can lead to cognitive biases and suboptimal choices. The study argues that individuals often fail to engage analytically (System 2 thinking) with AI-generated explanations due to the cognitive effort required, instead developing general heuristics about the AI's competence.

The authors propose and evaluate cognitive forcing functions as interventions to disrupt heuristic reasoning and promote analytical engagement with AI explanations. An experiment (N=199) compared three cognitive forcing designs with explainable AI approaches and a no-AI baseline. The findings indicate that cognitive forcing significantly reduced overreliance compared to simple explainable AI methods. However, a trade-off was observed: interventions that most effectively reduced overreliance received less favorable subjective ratings. Further analysis revealed that cognitive forcing interventions disproportionately benefited individuals with a higher Need for Cognition (NFC), suggesting that cognitive motivation moderates the effectiveness of explainable AI solutions. The research concludes that cognitive forcing functions hold potential for mitigating overreliance on AI, but their design must consider usability, acceptability, and potential intervention-generated inequalities related to cognitive motivation.

Theoretical Frameworks of Human Decision Making: A Bounded Rationality Perspective

This section examines the theoretical underpinnings of human decision-making, specifically focusing on the concept of bounded rationality and its implications for wishful thinking. Neumann, Rafferty, and Griffiths (2013) explore wishful thinking, the tendency to overestimate the probability of favorable outcomes and underestimate unfavorable ones, not as an irrational bias, but as a potentially adaptive strategy within the constraints of limited cognitive resources.

Traditional explanations of wishful thinking often characterize it as an irrational cognitive bias, potentially leading to risky choices. These explanations include ego-utility theory, which posits that wishful thinking protects self-image, and strategic theory, which suggests it can be tempered by incentives for accuracy. However, experimental evidence challenges these theories, demonstrating wishful thinking in situations unrelated to self-image and limited effects of accuracy incentives (Babad, 1997; Mayraz, 2011). Alternative explanations attribute wishful thinking to other cognitive biases like confirmation bias and cognitive dissonance (Kahneman, Slovic, & Tversky, 1982; Knox & Inkster, 1968).

Neumann et al. (2013) propose a contrasting perspective, framing wishful thinking as a heuristic that can improve decision-making when cognitive resources are limited. They model decision-making as a Markov Decision Process (MDP), a decision-theoretic model of sequential decision-making under uncertainty (Sutton & Barto, 1998). MDPs incorporate actions with uncertain effects and consider both immediate and future consequences. The core components of an MDP are states (S), actions (A), a transition model (T) defining the probability of transitioning between states given an action, a reward model (R) representing the immediate reward for each state-action pair, and a discount factor (b) weighting immediate versus future rewards. The optimal action is determined by calculating Q-values, representing the expected value of taking an action in a given state and timestep, considering both immediate and long-term rewards.

The authors simulate limited cognitive resources by restricting the planning horizon, i.e., the number of future time steps considered when making decisions. Solving for the optimal policy in an MDP scales exponentially with the planning horizon. By introducing a "wishful thinking" bias, operationalized as inflated probabilities of transitioning to high-value states, the model demonstrates that agents with moderate optimism can outperform unbiased agents when the planning horizon is limited. This suggests that wishful thinking can compensate for the inability to fully evaluate long-term consequences, leading to improved performance in certain decision-making environments.

Theoretical Frameworks of Human Decision Making

This course, "Behavioral Economics and Decision-Making," provides an overview of decision-making processes, encompassing psychological mechanisms, biases, and heuristics. It challenges classical economic theory by exploring cognitive processes involved in weighing options and making choices. The course emphasizes the influence of key behavioral economics principles, heuristics, and biases on decision-making.

The curriculum incorporates several theoretical models and concepts relevant to understanding decision-making. These include:

- **Libertarian Paternalism:** This framework is explored through lectures and debates, examining its theoretical underpinnings and practical implications.

- **Bounded Rationality:** The course addresses the concept of bounded rationality as a survival mechanism and its role in conserving cognitive resources.

- **Cognitive Dissonance:** The economic consequences of cognitive dissonance are examined, providing insights into how individuals reconcile conflicting beliefs and behaviors.

- **Behavior Change Models:** The course introduces and analyzes various behavior change models, including MINDSPACE, the EAST model, and the Com-B model, to understand how these frameworks can be applied to influence behavior.

- **EAST Model:** Specific elements of the EAST model (Easy, Attractive, Social, Timely) are discussed, with a focus on defaults, status quo bias, and cognitive overload. The course also covers biases such as availability bias, anchoring, loss aversion, and optimism bias, all within the context of the EAST model. Confirmation bias, commitment bias, and authority bias are also addressed.

- **Prospect Theory:** Prospect theory is examined through academic articles and video lectures, providing an analysis of decision-making under risk and uncertainty.

- **Group Polarization:** The phenomenon of group polarization is explored, drawing on research by Sunstein and Hastie, to understand how group dynamics can influence decision outcomes.

The course methodology combines theoretical instruction with practical application. Students are expected to identify and analyze real-life examples of these principles and biases, fostering a deeper understanding of their impact on decision-making in various contexts.

Theoretical Frameworks of Human Decision Making

Economic, psychological, and biological disciplines have developed formal models to understand decision-making processes under uncertainty, aiming to identify how decision-makers maximize value. These models serve as benchmarks to evaluate the optimality of decisions and refine understanding when behavior deviates from model predictions. However, these fields differ in their assumptions about optimal responses to uncertainty, particularly regarding what individuals aim to maximize. Economics and psychology often assume maximization of internal or subjective "goodness" (utility), while biological models prioritize long-term fitness benefits. This divergence leads to differing predictions about decision-making strategies. Furthermore, models vary in their purpose, with some prescribing ideal behavior (prescriptive models) and others describing actual behavior (descriptive models).

Classical economics, exemplified by Pascal's wager (1670/1995), posits that decisions under uncertainty should be based on the value of each outcome weighted by its likelihood, favoring the action with the greatest expected value. However, the St. Petersburg paradox, formalized by Bernoulli (1738), demonstrates that individuals do not always maximize expected value, instead weighting the utility of outcomes. Bernoulli argued that utility exhibits diminishing marginal returns, meaning that the impact of a reward diminishes as the initial reward size increases.

Von Neumann and Morgenstern (1947) formalized expected utility theory with four axioms defining rational decision-making: completeness (well-defined preferences), transitivity (consistent preference ranking), continuity (preferences exist on a common comparative scale), and independence (preferences unaffected by irrelevant alternatives).

Theoretical Frameworks of Human Decision Making

Psychological approaches have become increasingly prominent in the study of international politics, shifting the focus from structural and rationalist perspectives to the cognitive properties of actors themselves. This body of literature highlights that political elites, like ordinary citizens, are susceptible to misperceptions, motivated reasoning, and cognitive heuristics and biases. These biases, such as risk-seeking in the domain of losses, the intentionality bias, and reactive devaluation, can lead to more hawkish foreign policy decisions, increasing the risk of conflict.

Kahneman and Renshon (2007) argue that many cognitive biases identified by psychologists can lead political leaders to make more hawkish decisions. For example, the tendency to take risks to avoid a loss can encourage leaders to prolong wars, while biases in assessing adversaries' motives can escalate conflict. The intentionality bias, where actions with negative consequences are more likely to be deemed intentional, and reactive devaluation, the discounting of proposals from adversaries, can also worsen conflict.

However, a significant challenge exists in applying these individual-level psychological insights to group decision-making contexts, which are common in foreign policy. Critics argue that these biases may be mitigated or canceled out in group settings. Levy (1997) notes that concepts like loss aversion and framing, developed for individuals, may not automatically apply at the collective level. Hafner-Burton et al. (2017) express similar concerns, noting that institutional structures are often designed to mitigate individual biases.

The aggregation of psychological biases in groups remains an empirical question. Arrow's impossibility theorem demonstrates that even rational individuals in a group can produce irrational choices through aggregation mechanisms. Condorcet's jury theorem suggests that large groups can make better decisions if individuals vote independently and make the right choice with a probability greater than 50 percent; however, violating these assumptions can lead to worse decisions.

Kertzer, Holmes, LeVeck, and Wayne (2022) conducted three large-scale online experiments to test the aggregation of psychological biases in foreign policy. The experiments involved nearly 4,000 participants working through foreign policy scenarios individually or in groups with different structures. The findings indicate that biases such as risk-taking to avoid a loss, the intentionality bias, and reactive devaluation largely replicate in small-group contexts, with no significant reduction in group settings. In some cases, these biases may even be exacerbated.

Theoretical Frameworks of Human Decision Making

Decision making, viewed from a cognitive science perspective, is a high-level cognitive process involving choice among alternatives, influenced by perception, memory, attention, judgment, feedback, and learning. The field of "judgment and decision making" primarily focuses on the judgment process (evaluating alternatives) and the choice process (selecting among them).

Two dominant perspectives exist: the "open-loop" and "closed-loop" views. The open-loop model, prevalent in past decades, conceptualizes decision making as a linear process involving the presentation of choice options, beliefs about objective events (probabilities), and desires or utilities representing outcome consequences.

In contrast, the closed-loop view emphasizes an interactive and continuous dynamic process between humans and their environment. Decision makers perceive information, transform it to create alternatives, build preferences, evaluate options, execute actions, and process feedback to reinforce or adjust future decisions. This perspective aligns with models of dynamic decision making and decisions from experience in dynamic situations.

Historically, economic theories have assumed humans are utility maximizers ("rational"), while psychological research has demonstrated deviations from utility maximization ("irrational"). The principle of maximization of subjective expected utility (SEU) has been a central concept, but it has faced criticism regarding its applicability to human behavior. Herbert Simon proposed "bounded rationality" and the "satisficing" mechanism, suggesting that individuals adapt by selecting satisfactory actions without employing complex, cognitively demanding rules.

Theoretical Frameworks of Human Decision Making: A Categorical Model

This section examines a theoretical framework for human decision-making based on the principle that individuals process information using categories. Fryer and Jackson (2007) propose a model where decision-makers store past experiences in a finite set of categories, necessitating the grouping of heterogeneous experiences. Prototypes for prediction are then formed based on aggregated memory or statistics from each category. When faced with a new situation, the decision-maker identifies the most analogous category and makes predictions based on its prototype.

The research focuses on "optimal" categorizations, defined as those minimizing within-category variation, which, under certain conditions, is equivalent to maximizing expected utility. A key finding is that optimal categorization leads to less frequent experiences being lumped into more heterogeneous categories. This implies that interactions with minority groups, being less frequent for most decision-makers, are often sorted more coarsely than interactions with larger groups. The authors establish this through results that partially characterize optimal categorization, demonstrating how object categorization depends on relative frequencies and inter-object distance.

The authors acknowledge the difficulty in proving general results about optimal categorizations, highlighting the potential for unpredictable and volatile decision-making, even within individuals, as experiences change. The contribution is twofold: first, the development of a model linking categorization to biased decision-making; and second, the application of this model to specific social contexts. For example, in a labor market context, minorities may not be as finely sorted based on human capital investments, reducing their incentive to invest and perpetuating coarse sorting.

While acknowledging a rich literature on categorization, the authors emphasize the novelty of their model in providing provable results regarding the properties of optimal categorization, particularly its implication of differential treatment based on group size. This makes the model suitable for analyzing how categorization leads to specific and predictable biases in decision-making. The methodology involves constructing a formal model and deriving analytical results regarding the properties of optimal categorization. The conclusions suggest that categorization, while a necessary cognitive tool, can lead to biases, particularly against minority groups, even in the absence of explicit discriminatory intent.

Theoretical Frameworks of Human Decision Making: An Examination of Algorithmic Influence

Recent advancements in Artificial Intelligence (AI), fueled by the proliferation of large datasets and enhanced sensory technologies, are increasingly impacting various sectors, including transportation, healthcare, and employment. This necessitates a critical examination of the theoretical frameworks underpinning human decision-making, particularly in the context of AI-driven systems.

Impact of AI on Labor and Management: AI's influence on labor is multifaceted. While some research suggests AI and automation may displace workers in predictable physical activities and data processing roles (e.g., taxi and truck drivers), other studies indicate AI transforms the nature of work itself. Algorithmic management, exemplified by platforms like Uber, leverages data collection and behavioral economics to influence worker behavior, raising concerns about information asymmetry and potential manipulation. Furthermore, AI-driven management systems employ invasive methods, capturing and analyzing employee computer activities and sentiments, impacting the shift towards a 'gig economy' characterized by on-call work and predictive scheduling. The impact extends to professions like medical diagnostics, where machine learning (ML) algorithms can process vast datasets of cases, potentially surpassing human experience. However, concerns arise regarding 'automation bias' and the potential reduction in critical thinking and learning.

Bias and Inclusion in AI Systems: A significant concern revolves around the design and social implications of AI decision-making systems, particularly regarding bias and inclusion. Bias manifests in various forms, including statistical bias (difference between estimator's expected value and true parameter value) and selection bias (errors due to unequal sampling probabilities). AI-driven systems exhibit racial disparities in disciplinary actions, predictive policing, and healthcare datasets, which are historically skewed towards specific demographics. Historical examples, such as the Book of Life registry project in South Africa and the National Security Entry-Exit Registration System, highlight the potential for AI to perpetuate discriminatory practices. The integration of AI and algorithmic decision-support systems in law enforcement, including police body cameras, raises concerns about biased datasets and the need for transparency and accountability.

Rights and Liberties in the Age of AI: The deployment of AI systems impacts citizens' rights and liberties, particularly in areas like criminal justice and law enforcement. Risk assessment algorithms, while potentially improving accuracy in risk forecasting, are susceptible to limitations in data and features, leading to unwarranted trust. Privacy emerges as a critical concern, with asymmetries between institutions accumulating data and individuals generating it. The expansion of AI into urban planning, through the deployment of IoT devices and sensors, raises concerns about tracking human movements and preferences, potentially leading to predictive privacy harms.

Ethics and Governance of AI: The ethical considerations surrounding AI systems are paramount, particularly in sensitive or high-stakes contexts. Decision-making processes by AI systems and their developers are often obscured from public view and accountability. Ethical codes developed by organizations like the Future of Life Institute, IEEE, and ACM aim to set moral precedents and initiate conversations. However, the effective invisibility and pervasiveness of many AI systems hinder public discourse and the ability to opt-out. The development of ethically and societally relevant technologies based on algorithms, data, and AI (ADA) requires addressing the lack of clarity and consensus around central ethical concepts like privacy, bias, and explainability.

Tensions Between Values in AI Development: The development and deployment of AI systems involve inherent tensions between competing values. These tensions include: quality of services versus privacy, personalization versus solidarity, privacy versus transparency, convenience versus dignity, accuracy versus explainability, accuracy versus fairness, preferences versus equality, and efficiency versus safety. Identifying and resolving these tensions is crucial for ensuring that ADA-based technologies are developed and used for the benefit of society.

AI's Influence on Cognitive Processes

AI's Influence on Cognitive Processes: Algorithmic Bias Mitigation

Recent research highlights the vulnerability of machine learning systems to bias, particularly affecting under-represented segments of society (Amini et al., 2019). This bias can manifest in various AI applications, influencing decisions related to loan eligibility (Khandani, Kim, and Lo, 2010), criminal sentencing (Berk, Sorenson, and Barnes, 2016), news presentation (Nalisnick et al., 2016), and medical diagnoses (Mazurowski et al., 2008). The development of fair and unbiased AI systems is crucial to prevent unintended side effects and ensure the long-term acceptance of these algorithms (Miller, 2015; Courtland, 2018).

Algorithmic bias has been observed even in seemingly simple tasks like facial recognition (Zafeiriou, Zhang, and Zhang, 2015; Buolamwini and Gebru, 2018). For instance, Klare et al. (2012) found significantly lower accuracy in face detection systems used by US law enforcement when identifying dark-skinned women between 18 and 30 years old. This is particularly concerning as these systems are often integrated into larger surveillance or criminal detection pipelines (Abdullah et al., 2017).

To address this challenge, Amini et al. (2019) developed a novel, tunable algorithm for mitigating hidden biases within training data. Their approach fuses the original learning task with a variational autoencoder (VAE) to learn the latent structure within the dataset. The learned latent distributions are then adaptively used to re-weight the importance of certain data points during training. This method is generalizable across various data modalities and learning tasks. The algorithm identifies under-represented parts of the training data in an unsupervised manner and subsequently increases their respective sampling probability.

The effectiveness of this debiasing approach was evaluated on the Pilot Parliaments Benchmark (PPB) dataset (Buolamwini and Gebru, 2018), which is specifically designed to assess biases in computer vision systems. The results demonstrated increased overall performance and decreased categorical bias compared to standard deep learning classifiers.

The key contributions of Amini et al. (2019) include: (1) a novel debiasing algorithm that utilizes learned latent variables to adjust the sampling probabilities of individual data points during training; (2) a semi-supervised model for simultaneously learning a debiased classifier and the underlying latent variables governing the given classes; and (3) an analysis of their method for facial detection with biased training data, and evaluation on the PPB dataset to measure algorithmic fairness across race and gender. This approach differs from existing methods that rely on data pre-processing, artificially generated debiased data (Calmon et al., 2017), or resampling (More, 2016) by learning the latent structure of data using a VAE-based approach for resampling based on the data's underlying latent structure, debiasing automatically during training, and not requiring any data pre-processing or annotation prior to training or testing.

AI's Influence on Cognitive Processes: Algorithmic Bias and its Societal Impact

The advent of artificial intelligence (AI) has amplified the impact of biases inherent in decision-making processes, extending their reach beyond localized contexts. While human decision-making has historically been susceptible to bias, AI's widespread application introduces new challenges. Research indicates that AI algorithms can perpetuate and even exacerbate existing societal biases, leading to discriminatory outcomes across various domains.

Empirical Evidence of Algorithmic Bias:

Several studies have demonstrated the presence and consequences of algorithmic bias. For instance, research has revealed that:

- Mortgage pricing algorithms charge colored and Latino consumers more than White consumers for purchasing and refinancing mortgages.
- Corporate hiring algorithms exhibit bias in favor of men.
- Prison sentencing risk assessment algorithms are biased against African Americans.
- Facial recognition systems misinterpret the emotions of individuals based on their race.
- Facial recognition systems identify eyes of Asians as 'closed'.

These findings underscore the potential for AI to perpetuate discriminatory practices on a large scale.

Sources of Algorithmic Bias:

The root of algorithmic bias lies primarily in the data used to train AI models. Algorithms learn patterns from training data, and if this data reflects existing biases, the resulting AI system will likely replicate and amplify those biases. Key sources of bias include:

- **Biased Training Data:** If an AI algorithm is trained on historical data reflecting discriminatory practices (e.g., judicial judgments unfavorable to specific ethnic groups), it will likely replicate those biases.
- **Under-representation:** When certain demographic groups or decisions are under-represented in the training dataset, the resulting AI algorithm will exhibit bias against those groups. For example, speech recognition systems trained on datasets with limited voice samples from Asian women may struggle to accurately comprehend their speech.

Mitigation Strategies:

Researchers and practitioners have proposed several strategies to mitigate algorithmic bias, including:

- Using more representative training datasets.
- Eliminating features from the dataset that can be used to identify gender, age, race, etc.
- Modifying the dataset through feature engineering to make gender, age, race undifferentiated.
- Masking the features related to gender and race in the dataset by random shuffle.
- Data augmentation.

Societal and Legal Responses:

The recognition of algorithmic bias has spurred social activism and legislative action. Organizations such as Data for Black Lives, the Algorithmic Justice League, and AI for the People are actively working to raise awareness, conduct research, and advocate for policy changes. These efforts have contributed to the development of legislation aimed at promoting transparency, fairness, and accountability in algorithmic decision-making. Examples of such legislation include:

- The Equal Credit Opportunity Act (ECOA): Used to challenge AI algorithms that make discriminatory credit decisions.
- The General Data Protection Regulation (GDPR): Provides consumers with the right to obtain an explanation of automated decisions and challenge them.
- The Algorithmic Accountability Act: Requires companies to assess their automated decision systems for impacts on accuracy, fairness, bias, discrimination, privacy, and security.

Conclusion:

AI's influence on cognitive processes extends beyond mere automation and efficiency gains. The potential for algorithmic bias to perpetuate and amplify societal inequalities necessitates careful consideration of data quality, algorithm design, and ethical implications. Ongoing research, social activism, and legislative efforts are crucial for ensuring that AI systems are developed and deployed in a manner that promotes fairness, equity, and social justice.

AI's Influence on Cognitive Processes: Algorithmic Bias

Algorithmic bias, a phenomenon where machine learning programs inherit social patterns reflected in their training data, exhibits similarities to human cognitive bias, suggesting a shared origin in information processing. Johnson (2016) argues that this emergent nature obscures the bias, complicating identification and mitigation.

One key challenge is the reliance on proxy attributes: seemingly innocuous attributes correlated with socially-sensitive attributes. Attempts to mitigate bias by discouraging reliance on these proxies risk a trade-off with judgment accuracy, presenting a dilemma without a purely algorithmic solution.

Empirical evidence supports the existence of algorithmic bias. Caliskan et al. (2017) demonstrated that word-embedding machine learning, trained on a large internet text corpus, replicated human-like semantic biases. The software exhibited tendencies to associate stereotypical female names with family terms, African-American names with unpleasant words, and male names with science and math terms. These biases likely stem from patterns within the training data, reflecting existing social biases in online language use. The methodology involved training parsing software on a large dataset and then analyzing the resulting semantic associations produced by the program. The conclusions drawn were that machine learning algorithms can inadvertently learn and perpetuate human biases present in the data they are trained on.

AI's Influence on Cognitive Processes: Bias in AI Systems

This section explores the influence of AI systems on cognitive processes, specifically focusing on the manifestation and implications of bias within these systems. The analysis draws upon insights from a podcast featuring Carol Smith, a senior research scientist in human-machine interaction, and Jonathan Spring, a senior vulnerability researcher.

Manifestation of Bias in AI:

Bias in AI systems is primarily attributed to the data used for training and the individuals involved in system development. A key finding highlighted is that AI systems trained on datasets predominantly composed of a specific demographic (e.g., white males) exhibit a tendency to misidentify individuals from other demographic groups. This was exemplified by a 2019 NIST study that found commercial facial recognition applications frequently misidentified individuals of Asian or African American origin more often than white males.

Furthermore, the use of historical data, which often reflects existing societal biases, perpetuates discriminatory outcomes. For instance, AI systems trained on historical mortgage data may inadvertently discriminate against Latino and Black borrowers due to the embedded biases in past lending practices.

Challenges in Identifying and Mitigating Bias:

A significant challenge lies in the difficulty of detecting bias, particularly when sensitive information (e.g., race) is intentionally omitted from the dataset. While removing explicit identifiers might seem like a solution, correlated data such as ZIP codes can still reveal underlying patterns that lead to discriminatory outcomes. AI systems are adept at identifying patterns, even those unintended by the developers, which can inadvertently amplify existing biases.

Methodologies for Testing and Addressing Bias:

To proactively address bias, a more skeptical and thorough examination of data provenance is crucial. This involves understanding how and why the data was collected, as well as its intended purpose. The "Datasheets for Datasets" methodology is cited as a valuable resource for conducting in-depth data analysis. Additionally, speculative analysis, which involves anticipating potential consequences and worst-case scenarios, can help mitigate risks associated with combining multiple datasets or accessing sensitive data sources. This may involve considering alternative data sources or even questioning the suitability of AI for a particular application.

AI's Influence on Cognitive Processes: A Literature Review

This section reviews the influence of artificial intelligence (AI) on cognitive processes, focusing on psychological factors that shape attitudes toward AI tools. De Freitas, Agarwal, Schmitt, & Haslam (2023) identify five primary categories of resistance to AI: opacity, emotionlessness, rigidity, autonomy, and group membership, linking these barriers to fundamental aspects of cognition.

Opacity and Explainability: The perceived "black box" nature of AI systems can lead to distrust, as individuals are motivated to understand and predict their environment (De Freitas et al., 2023). Research indicates that individuals prefer AI systems that offer explanations of their decision-making processes (De Freitas et al., 2023). For instance, participants engaged more with advertisements for a skin cancer detection application when the algorithm's process was explained (e.g., checking mole shape, size, and color) compared to advertisements that only described the AI's function (De Freitas et al., 2023). However, the effectiveness of explanations varies; "why" explanations (e.g., braking due to an obstacle) are more effective than "how" explanations (e.g., the car is braking) in fostering positive attitudes (De Freitas et al., 2023). Contrastive explanations, which clarify why alternative outcomes were rejected, also enhance trust (De Freitas et al., 2023). Furthermore, the importance of explainability increases with the stakes of the decision (De Freitas et al., 2023).

Illusory Transparency of Human Decision-Making: The preference for human decision-making over AI often stems from a perceived transparency that is, in many cases, illusory (De Freitas et al., 2023). Individuals tend to overestimate their understanding of human decision-making processes, relying on heuristics rather than direct insight (De Freitas et al., 2023). Interventions that highlight the opacity of human decision-making can improve attitudes toward AI (De Freitas et al., 2023). A study involving medical diagnoses found that when participants were asked to explain human and AI solutions, they experienced a greater reduction in their subjective understanding of human decisions, thereby improving their perception of AI (De Freitas et al., 2023).

Complexity and Task Appropriateness: The acceptance of explainable AI is contingent on the perceived appropriateness of its complexity for the task at hand (De Freitas et al., 2023). If explanations suggest that an AI system is either too simple or too complex for a given task, users may be less likely to trust its recommendations (De Freitas et al., 2023).

AI's Influence on Cognitive Processes: Overreliance and the Role of Explanations

Overreliance on AI systems, defined as the tendency to accept AI predictions even when incorrect, poses a significant challenge to effective human-AI collaboration. While Explainable AI (XAI) has been proposed as a solution, previous research has yielded limited evidence supporting its ability to mitigate overreliance. Some studies suggest that explanations may even exacerbate the issue by increasing trust or anchoring users to the AI's prediction. However, this paper challenges the notion that overreliance is an inevitable consequence of human cognition, arguing instead that individuals strategically choose whether to engage with AI explanations based on a cost-benefit analysis.

This research posits that the cognitive costs associated with both the task itself and the AI's explanation play a crucial role in determining the extent of overreliance. The authors propose that prior studies may have failed to demonstrate the effectiveness of explanations due to focusing on scenarios where the cognitive costs of verifying the explanation are comparable to the costs of completing the task independently. In such situations, individuals may opt to rely solely on the AI's prediction to minimize cognitive effort.

To investigate this hypothesis, the authors developed a cost-benefit framework that models the strategic decision-making process involved in human-AI interaction. This framework suggests that explanations will reduce overreliance when the task is sufficiently challenging, and the explanations significantly reduce the cost of verifying the AI's prediction, or when the explanations are inherently easier to understand.

This framework was empirically tested through five studies (N=731) utilizing a maze-solving task. Study 1 manipulated task difficulty, demonstrating that explanations reduced overreliance only when the task was sufficiently effortful. Studies 2 and 3 focused on the cognitive cost of understanding the AI's explanations, finding that lower cognitive effort led to a greater reduction in overreliance. Study 4 examined the interplay between cognitive effort and incentives, suggesting that individuals weigh the costs of engaging with explanations against the potential benefits. Finally, Study 5 employed the Cognitive Effort Discounting paradigm to quantify the utility of different explanations, providing further support for the proposed cost-benefit framework.

The findings of this research suggest that the effectiveness of AI explanations in reducing overreliance is contingent upon the relative cognitive costs associated with the task and the explanation. The results indicate that the null effects observed in previous literature may be attributable to explanations not sufficiently reducing the cognitive burden of verifying the AI's prediction.

AI's Influence on Cognitive Processes: A Literature Review

The detrimental effects of bias in Artificial Intelligence (AI) on societal equality have been documented, with predictive policing systems, healthcare algorithms, and biometric technologies exhibiting discriminatory outcomes against marginalized groups (Richardson et al., 2019; Obermeyer et al., 2019; Buolamwini and Gebru, 2018; Datta et al., 2015; Lambrecht and Tucker, 2019; Noble, 2018). These biases often stem from training data that reflects historical injustices and societal inequalities (Perez, 2019; Hickman, 1997).

Current technical approaches to mitigating bias in AI training data often aim to re-balance data and achieve neutrality concerning protected classes, employing methods such as data augmentation, re-sampling, and re-weighting (Kamiran and Calders, 2012; Feldman et al., 2015). In natural language processing, techniques like word embedding models are used to identify and address social and intersectional biases in text (Tan and Celis, 2019), with interventions involving disassociating stereotypical relationships and swapping gendered entities (Bolukbasi et al., 2016; Zhao et al., 2018; Park et al., 2018). However, evaluations of bias in these studies often focus on specific aspects, such as stereotypical associations in employment, or rely on implicit association tests (Caliskan, 2017).

Despite efforts to achieve objectivity, there is a growing recognition that data-driven AI inherently involves bias (Meredith, 2018; O'Neil, 2016; Dignum, 2019; Bartoletti, 2020). Data collection and curation for machine learning algorithms are viewed as political acts that perpetuate specific worldviews (Noble, 2018; Gebru, 2019). A critical framework for assessing the ideology underlying datasets, particularly concerning gender representation, emphasizes that data are not neutral or objective (D'Ignazio and Klein, 2020). This perspective suggests a shift from merely fixing bias to curating data collections that intentionally reflect desired representations of race, gender, and social class.

Furthermore, the collection of more data from under-represented groups to address bias raises ethical concerns regarding data gathering practices (Hawkins, 2018; Benjamin, 2019; Eubanks, 2018). The socio-political ramifications of increased identifiability within unjust social structures and the historical legacy of inequitable treatment in data collection must be considered. The role of data curators in classifying and labeling AI training data is highlighted as central to preventing discrimination in AI.

AI's Influence on Cognitive Processes: Overreliance and Cognitive Forcing

Recent research indicates that human-AI teams, while generally outperforming individuals, often underperform compared to the AI alone due to overreliance on AI suggestions, even when incorrect (Buçinca, Malaya, & Gajos, 2021). Explainable AI (XAI), intended to mitigate this by providing rationales for AI recommendations, has not substantially reduced this overreliance (Buçinca et al., 2021).

Drawing on dual-process theory, Buçinca et al. (2021) posit that individuals primarily engage in System 1 thinking (heuristic-based) rather than analytical System 2 thinking when interacting with AI explanations. This leads to the development of general heuristics about AI competence rather than individual evaluation of each explanation's content.

To address this, the study investigated cognitive forcing functions – interventions designed to disrupt heuristic reasoning and promote analytical engagement. An experiment (N=199) compared three cognitive forcing designs with two simple XAI approaches and a no-AI baseline (Buçinca et al., 2021). The methodology involved presenting participants with AI-assisted decision-making tasks under different intervention conditions.

The results demonstrated that cognitive forcing significantly reduced overreliance compared to simple XAI approaches, although it did not eliminate it entirely (Buçinca et al., 2021). A trade-off was observed: interventions that most effectively reduced overreliance were also rated as less usable and acceptable. Furthermore, the study revealed that cognitive forcing interventions disproportionately benefited individuals with a higher Need for Cognition (NFC), suggesting that cognitive motivation moderates the effectiveness of XAI solutions (Buçinca et al., 2021).

AI's Influence on Cognitive Processes: A Literature Review

This section explores the influence of Artificial Intelligence (AI) on cognitive processes, drawing upon insights from Dede, Etemadi, and Forshaw (Next Level Lab, Harvard Graduate School of Education). The central argument posits that AI's increasing proficiency in calculation, computation, and prediction ("reckoning") necessitates a shift in workforce development towards uniquely human judgment skills, such as decision-making under uncertainty, deliberation, ethics, and practical knowing.

The authors define intelligence as "the disposition to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience." This definition emphasizes performance-based intelligence, focusing on the application of abilities rather than inherent capabilities. Furthermore, it incorporates a dispositional aspect, highlighting the importance of sensitivity and inclination in deploying cognitive capacities.

Within a workplace context, intelligence is divided into two complementary roles: reckoning and judgment. Reckoning encompasses calculative prediction and formulaic decision-making, areas where AI systems excel. Judgment, conversely, involves deliberative thought grounded in ethical commitment and situational appropriateness. The authors argue for a complementary partnership between AI and humans, where AI handles reckoning tasks with speed and scale, while humans contribute judgment based on contextual understanding and factors AI cannot readily incorporate. This synergistic combination, termed "intelligence augmentation," results in performance exceeding the sum of individual capabilities.

The paper distinguishes judgment and decision-making from soft skills, characterizing them as a unique category that leverages information from soft skills, contextual factors, and non-human information to determine a course of action. The authors emphasize that effective judgment requires strong soft skills but is not equivalent to them.

The analysis of future occupational skills, based on a 2017 study by Pearson and NESTA, suggests a need for workforce development programs to prioritize judgment skills to prepare individuals for a future where AI rapidly advances in reckoning capabilities. Hypothetical cases involving small business owners in food services and green energy illustrate the importance of developing these skills to maintain economic inclusion in a rapidly changing landscape.

AI's Influence on Cognitive Processes: A Literature Review

This section examines the influence of generative artificial intelligence (GAI) tools on cognitive processes within academic research, focusing on the potential for cognitive bias perpetuation. A study employing simulations across three GAI platforms (ChatGPT, Bard, and Copilot) investigated the propagation of misinformation and miscitations, particularly concerning common method variance (CMV) literature. The methodology involved presenting initial and follow-up prompts related to CMV, specifically targeting Harman's One-Factor Test, and comparing the generated results with established research.

The research findings indicate that GAI tools can disseminate inaccurate information and misattribute sources. Specifically, the simulations revealed that GAI platforms may perpetuate the use of Harman's One-Factor Test despite its documented unreliability in detecting CMV issues (Fuller et al., 2016). This perpetuation is attributed to a "bandwagon effect," where the widespread adoption of a technique, even if flawed, reinforces its perceived validity (Barrera & Ponce, 2021; Kessous & Valette-Florence, 2019). The GAI tools, in some instances, failed to adequately acknowledge the test's limitations, such as its low sensitivity and potential for false negatives (Lindell & Whitney, 2001), and the fact that it is deficient in detecting CMV (Baumgartner et al., 2021).

The study concludes that GAI tools should not be blindly trusted in academic research. Instead, researchers are advised to adopt an "augmentation" approach, utilizing AI in close collaboration with human expertise (Raisch & Krakowski, 2021). Recommendations include rigorous verification of GAI-generated results, iterative prompting to refine outputs, meticulous examination of in-text citations and bibliographies, and direct consultation of original sources (De Lacey, Record, & Wade, 1985). While acknowledging the potential value of GAI for novice researchers, the study emphasizes the importance of mitigating cognitive biases and ensuring the reliability of research methods to maintain rigor in academic publications.

Ethical Considerations of AI-Driven Decisions

Ethical Considerations of AI-Driven Decisions: A Deontological Approach

This section reviews the ethical considerations of AI-driven decisions, focusing on the limitations of intuition-based approaches and advocating for a deontological framework. Anderson and Anderson (A&A) have significantly contributed to machine ethics by advocating for principle-based systems with non-consequentialist elements to respect individual dignity (Anderson & Anderson, 2007, 2010, 2011, 2014, 2015a, 2015b; with Armen, 2006; with Berenz, 2017). A&A's approach, which computerizes W.D. Ross's prima facie duty framework, is critiqued for its reliance on human moral intuition and its potentially inadequate treatment of autonomy.

A&A's methodology involves inductive logic programming to derive decision principles within specific domains, such as healthcare. For instance, in a scenario where a patient refuses medication, the system evaluates options (Notify or Don't Notify) based on the degree to which they satisfy prima facie duties like non-maleficence, beneficence, and autonomy, drawing upon the work of biomedical ethicists Buchanan and Brock (1990) and Beauchamp and Childress (1979). The system assigns numerical values to these duties and seeks an equilibrium that aligns with ethicists' intuitions regarding the correct course of action.

However, the reliance on moral intuition is identified as a significant weakness. Moral intuitions are susceptible to situational cues and may not be consistent (Alexander, 2012; Appiah, 2008; Sinnott-Armstrong, 2006). Furthermore, research suggests that professional ethicists' intuitions are not significantly different from those of ordinary individuals (Schwitzgebel & Rust, 2016). The paper argues that a core motivation for developing machine ethics is to minimize human error, including reliance on potentially unreliable human moral intuitions. A&A's system also struggles in scenarios where ethicists lack consensus, limiting its ability to address novel ethical dilemmas.

The paper also critiques A&A's treatment of autonomy. While A&A incorporate deontological elements to protect individual dignity, their approach may not adequately address situations where violating one person's autonomy maximizes benefits for a larger group. The paper posits that A&A's system might require assigning disproportionately high negative values to autonomy violations to prevent ethically questionable outcomes, highlighting a potential inconsistency in their framework.

Ethical Considerations of AI-Driven Decisions: A Literature Review

Artificial intelligence (AI) is fundamentally transforming human interactions and societal systems, introducing risks such as reduced human autonomy, algorithmic bias, data privacy threats, and accountability challenges for algorithmic harm. Emerging technologies like embodied AI and large language models exacerbate these risks, affecting human-machine interactions and raising concerns about environmental sustainability and human rights across the AI value chain. Policymakers are increasingly adopting a human rights lens for AI governance in response to these societal impacts.

Current regulatory frameworks often fail to address the global digital divide, the disproportionate risks AI poses to marginalized communities, and human rights issues embedded in the AI value chain. The impact of AI extends beyond traditional human rights, intersecting with categories of international law, such as the militarization of AI and its alignment with international humanitarian law.

In response to these challenges, the Munich Convention on AI, Data, and Human Rights was drafted, aiming to advance global AI governance based on human rights principles within international law. This framework calls for the prohibition of harmful AI uses, the definition of domains critical to the enjoyment of AI, and the establishment of robust enforcement mechanisms to prevent adverse structural human rights impacts created by or linked to AI.

The UN Charter and the Universal Declaration of Human Rights form a cornerstone of international law, addressing the diverse needs of society, including marginalized groups, while responding to emerging technological, humanitarian, political, and social challenges. The interconnected nature of AI calls for a legal discourse that considers diverse perspectives and transcends traditional branches of international law.

Ethical Considerations of AI-Driven Decisions

The increasing prevalence of Artificial Intelligence (AI) and related technologies presents significant ethical challenges, particularly concerning potential harms such as compromised privacy, security, and safety, as well as the perpetuation of bias, diminished accountability, and workforce displacement. Addressing these challenges necessitates an interdisciplinary approach, integrating expertise from law, ethics, policy, business, engineering, and computer science. Organizations are increasingly adopting "ethics by design" teams to proactively address ethical and legal compliance throughout the product lifecycle, from design to retirement.

Research initiatives are actively investigating algorithmic bias, discrimination, transparency, and distributional concerns in AI development and deployment. The Science, Law, and Policy (SLAP) Lab utilizes novel data to address legal and policy questions related to cutting-edge science and technology. Furthermore, research is being conducted on developing adaptable ethics and compliance frameworks for data commons, drawing upon legal and policy developments such as the EU Data Governance Act. The exploration of standardized form agreements for data licensing, similar to open-source models, is also underway.

Ethical Considerations of AI-Driven Decisions: A Literature Review

This section examines the ethical considerations surrounding the increasing use of Artificial Intelligence (AI) in decision-making processes that significantly impact individuals' lives. It draws upon the principles of algorithmic transparency and accountability as outlined by the Association for Computing Machinery (ACM) and identifies common pitfalls in deployed AI systems.

Principles for Transparency and Accountability:

The ACM's United States and European Public Policy Councils (ACM Public Policy Council 2017; Garfinkel et al. 2017) have articulated seven key principles to enhance human oversight of AI systems:

- **Awareness:** Stakeholders must recognize potential biases in AI design, implementation, and use, and understand the potential harm these biases can inflict.
- **Access and Redress:** Mechanisms should be in place to allow individuals and groups adversely affected by algorithmic decisions to question and seek redress.

- **Accountability:** Institutions must be held responsible for decisions made by the algorithms they employ, even when the algorithms' processes are opaque.
- **Explanation:** Systems should provide explanations of both the procedures followed by the algorithm and the specific decisions made.
- **Data Provenance:** The origins of training data should be documented, along with an analysis of potential biases introduced during data collection.
- **Auditability:** Models, algorithms, data, and decisions should be recorded to facilitate auditing in cases of suspected harm.
- **Validation and Testing:** Rigorous methods should be used to validate models, and routine tests should be conducted to assess and identify discriminatory harm.

Antipatterns in Deployed AI Systems:

Failures to adhere to these principles often result in harmful trends or "antipatterns." These include:

- **Learning from the Past without Context:** AI systems trained on historical data can perpetuate and even amplify existing societal biases, such as gender imbalances in specific professions. Even when protected attributes like race and gender are removed, proxy variables (e.g., zip code, magazine subscriptions) can still encode these biases (Duhigg 2012; U.S. Consumer Financial Protection Bureau 2014; Allen 2018; Rivera and Tilcsik 2016; Dastin 2018). Zhao et al. (2017) demonstrated that models trained on image datasets can amplify existing gender disparities. Caruana (2017) suggests retaining bias features during training and removing learned biases afterward, although this may conflict with data protection regulations.
- **Making Spurious Correlations:** Machine learning systems can identify correlations that are not causally related to the desired outcome. Ribeiro et al. (2016) illustrated this by training a classifier to distinguish between dogs and wolves based on the presence of snow, highlighting the importance of explainable AI techniques.
- **Learning from Humans without Considering Malicious Training:** AI systems that learn directly from human input are vulnerable to absorbing both positive and negative behaviors. The Microsoft Tay chatbot incident (Lee 2016) exemplifies how adversarial attacks can lead to the generation of inappropriate content. Suciu et al. (2018) provide an overview of research on poisoning machine learning.

Ethical Considerations of AI-Driven Decisions: A Literature Review

This section draws upon a survey and legal analysis of existing United States law and regulations pertaining to AI and robotics, conducted as part of the SIENNA project (Bavitz, Holland, & Nishi, 2018). The analysis reveals a significant gap in cohesive federal policy concerning the development, implementation, and regulation of AI technologies within the U.S. legal framework.

Key Findings:

- **Lack of Cohesive Federal Policy:** The research highlights the absence of a unified federal approach to AI and robotics governance. This absence extends to constitutional amendments, human rights legislation, and the establishment of national regulatory bodies specifically dedicated to AI oversight (Bavitz, Holland, & Nishi, 2018).
- **Algorithmic Unfairness:** Algorithmic unfairness in decision-making processes is identified as a well-documented issue in the U.S. However, legal protections remain limited, with credit scoring representing one of the few areas with relatively more robust, albeit still incomplete, safeguards (Bavitz, Holland, & Nishi, 2018).
- **Private Sector Influence:** The study emphasizes the substantial influence of private industry standards and guidelines in shaping the development and regulation (or lack

thereof) of AI technologies. This influence underscores the need for greater public sector involvement in establishing ethical and legal frameworks (Bavitz, Holland, & Nishi, 2018).

Methodology:

The research employed a comprehensive survey and legal analysis of existing U.S. laws and regulations at the federal, state, and local levels. This included an examination of "hard" law, "soft" law, and relevant case law limited to federal and Supreme Court decisions. The objective was to assess the extent to which existing standards address the challenges posed by AI and robotics, identify gaps, and evaluate the validity, relevance, and adequacy of current regulations (Bavitz, Holland, & Nishi, 2018).

Conclusions:

The findings suggest that the ethical considerations of AI-driven decisions in the U.S. are currently addressed in a fragmented and incomplete manner. The lack of a cohesive federal policy, coupled with the significant influence of private sector standards, raises concerns about potential biases, unfairness, and a lack of accountability in AI systems. Further research and policy development are needed to establish robust ethical and legal frameworks that ensure fairness, transparency, and accountability in the design, deployment, and use of AI technologies.

Reference:

Bavitz, C., Holland, A., & Nishi, A. (2018). *Ethics and Governance of AI and Robotics: A Survey and Legal Analysis of Existing United States Law and Regulation*. Berkman Klein Center for Internet & Society.

Ethical Considerations of AI-Driven Decisions

The rapid advancement and implementation of artificial intelligence (AI) technologies carry significant ethical consequences, necessitating efficient governance to optimize benefits while minimizing risks. This governance must be based on sound values, including transparency, truth, safety, security, ethics, and privacy.

Transparency: Transparent AI systems foster trust, accountability, and fairness by allowing stakeholders to understand and evaluate decision-making processes, identify biases, and ensure ethical and legal compliance.

Truth: Trustworthy AI must produce accurate and unbiased data, preventing misinformation and promoting critical thinking, especially in sensitive areas like media, education, and science.

Safety and Security: Given AI systems' profound interactions with people, safety is vital to protect individuals and communities, reduce accidents and errors in AI-driven industries, and encourage ethical and legal compliance.

Ethics: Ethical considerations, including justice, transparency, accountability, privacy, and human rights, are needed to design and deploy AI technologies that benefit society and minimize harm.

Privacy: AI must respect privacy to secure personal data and retain trust in technology, protecting personal data from unlawful access and exploitation.

The proposed AI governance model consists of data feeding into algorithms, algorithms feeding into computing, and computing feeding into applications. Each of these areas requires specific governance:

Data Governance: AI data must be governed for ethical, accurate, and secure use, overseeing data collection, storage, analysis, and sharing while preserving sensitive data and preventing misuse.

AI Algorithmic Governance: AI algorithms must be governed to ensure ethical behavior and impartial results, preventing algorithmic biases and discrimination. The selection, design, training, and testing of algorithms must be governed.

Computing Governance: Computing technologies must be governed to ensure safe, ethical, and valuable AI technology, respecting privacy, improving security, and preventing harm. This includes governing semiconductor chips, edge computing, cloud computing, ambient computing, quantum computing, computing energy, and computing water use.

Governance of AI Applications: AI must be governed across society, economy, and politics to handle its consequences and potential disruptions, improving social well-being, economic fairness, and democratic processes without damaging them.

The document recommends aligning AI values with human rights and the United Nations Charter to ensure AI development supports human dignity, equality, and freedom. It also suggests utilizing behavioral science to cultivate a culture that optimizes AI's benefits while reducing risks, and implementing strategies based on mechanism design to foster a culture that maximizes the potential benefits of AI.

Ethical Considerations of AI-Driven Decisions: Literature Review

This section draws upon Ali et al.'s (2023) qualitative study to examine the ethical considerations surrounding AI-driven decisions within technology companies. The research addresses the gap in understanding the practical implementation of AI ethics, moving beyond theoretical discussions and critiques of "ethics washing."

Key Findings:

- **Decoupling of Ethics:** The study reveals a significant disconnect between publicly stated ethical principles and the actual practices and outcomes within technology companies. This "decoupling" phenomenon highlights the challenges in translating ethical values into tangible actions.
- **Role of Ethics Entrepreneurs:** AI ethics workers are identified as "ethics entrepreneurs" who attempt to institutionalize ethical practices within their organizations. These individuals face significant barriers in their efforts.
- **Prioritization Challenges:** Ethics-related concerns often struggle to gain priority within organizations primarily focused on rapid software product launches. Ethical considerations that might delay or impede launches are at risk of being overlooked.
- **Quantification Difficulties:** The study highlights the difficulty in quantifying the impact of ethics interventions, particularly in environments where employee incentives are tied to measurable financial gains from product innovation.
- **Organizational Instability:** Frequent reorganizations ("reorgs") within companies hinder the ability of ethics entrepreneurs to access institutional knowledge and maintain crucial relationships, thereby undermining their effectiveness.
- **Individualized Risk:** The research demonstrates that raising ethical concerns often places significant personal risk on individual employees, particularly those from marginalized

backgrounds. This contradicts the stated goals of promoting fairness and equity within the field of AI ethics.

Methodology:

The study employs a qualitative approach, including:

- **Interviews:** Twenty-five interviews were conducted with AI ethics workers within technology companies.
- **Observations:** Observations were made during industry workshops and training programs related to responsible AI.

Conclusions:

Ali et al.'s (2023) research underscores the complex challenges in implementing AI ethics within technology companies. The findings suggest that despite the growing emphasis on ethical AI, significant systemic barriers impede the effective integration of ethical considerations into AI-driven decision-making processes. The individualization of risk associated with raising ethical concerns further exacerbates these challenges, particularly for marginalized employees. The study calls for a deeper examination of organizational structures and incentive systems to facilitate a more robust and equitable implementation of AI ethics.

Ethical Considerations in AI-Driven Decisions: A Literature Review

The increasing prevalence of Artificial Intelligence (AI) and Autonomous Systems (AS) necessitates a thorough examination of the ethical norms, rules, and regulations governing their operation, as well as the ethical standards expected of designers and developers (Vakkuri, Kemell, & Abrahamsson, 2019). Developers embed their own values into systems, reflecting their perspectives, yet they often lack sufficient ethical awareness and tools to address potential ethical issues during development (Vakkuri et al., 2019).

One approach to address this gap is the RESOLVEDD strategy, a nine-step method designed to support users in considering and tackling ethical issues based on their values or ethical theories (Vakkuri et al., 2019). This strategy facilitates justification and explanation of the decision-making process leading to specific actions. The steps involve defining the background, conflict, possible solutions, consequences, impact, values upheld and violated, evaluation of outcomes, and justification of the chosen solution against objections (Vakkuri et al., 2019).

Another framework, Ethically Aligned Design (EAD), emphasizes the integration of ethical considerations into the design and decision-making processes of AI and autonomous systems (Vakkuri et al., 2019). This framework proposes AI-specific values and incorporates the ART principles: Accountability, Responsibility, and Transparency (Vakkuri et al., 2019). These principles provide a framework for ethical decision-making, promoting trust, integrity, and positive social impact (Vakkuri et al., 2019). Accountability involves explaining and justifying decisions to stakeholders, responsibility connects the designer to external stakeholders, and transparency focuses on the algorithms, data, and their dynamics (Vakkuri et al., 2019). Transparency is considered crucial for ethical AI design, enabling understanding of AI/AS actions and their rationale (Vakkuri et al., 2019).

Empirical findings suggest that while EAD can be introduced, its long-term sustainability requires adjustments. Group discussions were identified as a valuable means to implement EAD in practice (Vakkuri et al., 2019). Furthermore, the study revealed that the presence of an ethical tool, like RESOLVEDD, increases ethical consideration and responsibility among development teams (Vakkuri et al., 2019). However, questions of accountability in design remain complex, with teams sometimes disclaiming accountability, particularly for conceptual systems (Vakkuri et al., 2019). While RESOLVEDD promotes transparency in the design process, it does not necessarily translate to transparency at the product layer (Vakkuri et al., 2019).

References

Vakkuri, V., Kemell, K.-K., & Abrahamsson, P. (2019). *Ethically Aligned Design: An empirical evaluation of the RESOLVEDD-strategy in Software and Systems development context*. Faculty of Information Technology, University of Jyväskylä, Finland.

Ethical Considerations of AI-Driven Decisions

This section reviews the ethical considerations surrounding AI-driven decisions, focusing on fairness, accountability, and transparency. These three pillars are critical for responsible AI development and deployment.

Fairness: Fairness in AI is broadly defined as the absence of prejudice or preference based on individual or group characteristics. Data bias and model bias are key challenges to achieving fairness. Data bias arises during data collection and includes historical bias (pre-existing societal biases reflected in data, e.g., gender-biased analogies generated by models trained on Wikipedia), representation bias (skewed datasets due to sampling methods, e.g., facial recognition systems trained primarily on white faces), and measurement bias (inaccurate or noisy data used as proxies for actual features, e.g., poorly trained survey interviewers leading to inaccurate mortality rate estimations). Model bias, on the other hand, stems from the methods used to evaluate and make decisions. Evaluation bias occurs during model iteration when benchmarks do not represent the general population, and aggregation bias arises when distinct populations are inappropriately combined (e.g., NLP models falsely flagging text due to context-specific language).

Accountability: Accountability ensures that failures within AI systems can be corrected to mitigate harm. Key components include good governance, understanding data, defining performance goals and metrics, and monitoring performance. Good governance operates at both organizational and system levels. Organizational-level governance involves defining clear goals, roles, responsibilities, values, workforce composition, stakeholder involvement, and risk management. System-level governance emphasizes technical specifications, compliance with regulations, and transparency. Understanding data involves assessing the sources, reliability, categorization, variable selection, and enhancement methods for data used in model development and evaluating the dependency, bias, security, and privacy of data used in system operation. Defining performance goals and metrics requires documentation, precise metrics, and assessment at both component and system levels, including bias identification and human supervision. Continuous monitoring of performance and assessing sustainment and expanded use are crucial for maintaining accountability over time.

Transparency: Transparency refers to the ability to describe, inspect, and reproduce the mechanisms through which AI systems make decisions and adapt. Key components include traceability, communication, and intelligibility/explainability. These elements are essential for building trust and enabling effective oversight of AI systems.

Ethical Considerations of AI-Driven Decisions

Recent research highlights the vulnerability of machine learning systems to bias, particularly affecting under-represented segments of society. This paper addresses this issue by developing a novel algorithm to mitigate hidden biases within training data. The algorithm combines the original learning task with a variational autoencoder (VAE) to learn the latent structure within the dataset. It then adaptively re-weights data points based on the learned latent distributions.

The methodology involves an end-to-end deep learning algorithm that simultaneously learns the desired task (e.g., facial detection) and the underlying latent structure of the training data in an unsupervised manner. This enables the discovery of hidden biases. The algorithm identifies under-represented examples and increases their sampling probability during training.

The algorithm was evaluated on the Pilot Parliaments Benchmark (PPB) dataset to assess racial and gender bias in facial detection systems. Results demonstrated increased overall performance and decreased categorical bias compared to standard deep learning classifiers.

Key contributions include: (1) a tunable debiasing algorithm using learned latent variables to adjust data point sampling probabilities; (2) a semi-supervised model for learning a debiased classifier and the underlying latent variables; and (3) analysis of the method for facial detection with biased data, evaluated on the PPB dataset to measure algorithmic fairness across race and gender.

This research contrasts with existing approaches that focus on data pre-processing, in-processing, or post-processing for fairness. Many of these rely on artificially generated debiased data or resampling techniques that primarily address class imbalances rather than variability within a class. Unlike these methods, this approach uses a VAE-based method for resampling based on the data's underlying latent structure, debiasing automatically during training without requiring data pre-processing or annotation.

Methodology

Research Design and Data Collection Methods

Methodology

This research employs a case study methodology to investigate the impact of automatically generated explanations on user perceptions of Explainable AI (XAI) systems. The study focuses on how different styles of AI-generated rationales are perceived by non-expert users across dimensions of confidence, human-likeness, adequate justification, and understandability. The research design incorporates a data collection pipeline to create a corpus of natural language "think-aloud" explanation data, followed by two user studies that evaluate human factors.

Research Design and Data Collection Methods:

The study utilizes a modified version of the Frogger game as a sequential decision-making task environment. Participants simultaneously play the game and provide explanations for each action. The data collection process consists of three phases:

1. **Guided Tutorial:** Ensures user familiarity with the interface.
2. **Rationale Collection:** Participants engage in a turn-taking experience, explaining each action while the game is paused. An automatic speech-to-text library transcribes utterances to reduce participant burden.
3. **Transcribed Explanation Review:** Participants review all action-explanation pairs in a global context by replaying each action.

This process yielded over 2000 samples of action-explanation pairs from 60 participants, collected via Turk Prime.

Data Analysis:

The collected data was used to train an encoder-decoder recurrent neural network to generate natural language explanations for given actions. The study then analyzed the impact of these generated rationales on human-factor dimensions, such as confidence in the agent's decision, understandability, human-likeness, explanatory power, tolerance to failure and unexpected behavior, likeability, and perceived intelligence.

Methodology

This research article employs a literature review methodology to investigate the impact of Artificial Intelligence (AI) on decision-making processes within organizations. The study synthesizes existing academic literature to explore the consequences of AI implementation for individuals, companies, and society. The research design focuses on identifying and analyzing key themes related to AI's influence on internal operations across various departments, including transport and consultation management. Data collection involves gathering information from scholarly articles, industry reports, and other relevant sources to provide an overview of AI mechanisms such as automation efficiency, complex management, succession planning incorporation, and business decision support in risk monitoring, assessment, compliance, and scam recognition. The analysis encompasses the optimistic impacts of AI on decision-making, as well as potential challenges, including ethical concerns, algorithmic biases, and social allegations. The review also considers the importance of data privacy protection, transparency, accountability, and explainability in AI-driven decision-making.

Methodology

This research employed an experimental design to investigate the efficacy of cognitive forcing functions in mitigating overreliance on AI in decision-making contexts. The study compared three cognitive forcing designs with two explainable AI (XAI) approaches and a no-AI baseline, utilizing a sample of 199 participants recruited through Amazon Mechanical Turk. Data collection involved observing participant decision-making behavior under each condition, specifically measuring the extent to which participants followed AI suggestions, even when incorrect. Subjective ratings of the designs were also collected to assess usability and acceptability. Furthermore, the study incorporated the Need for Cognition (NFC) scale to evaluate individual differences in cognitive motivation and to audit for potential intervention-generated inequalities. The analysis focused on comparing overreliance rates across conditions and examining the relationship between NFC levels and the effectiveness of cognitive forcing interventions.

Methodology

This research employs a qualitative, multiple-case study design to investigate the impact of Artificial Intelligence (AI) on design practices. The study focuses on two pioneering organizations at the forefront of AI and design thinking: Netflix and Airbnb. Data collection is complemented by analyses of Microsoft and Tesla to provide further insights. The case studies offer a privileged perspective on the evolving nature of design in the age of AI. Data analysis involves examining the design practices and principles within these organizations to understand how AI is transforming the design process, the objects of design, and the fundamental principles that guide design activities.

The research aims to compare human-intensive design approaches with design practices in AI-driven environments.

Methodology

This research employs offline reinforcement learning (RL) to model human-AI decision-making and optimize human-AI interaction for diverse objectives, specifically human-AI accuracy and human skill improvement. The approach casts the problem of human-AI decision-making as a Markov Decision Process, learning policies from existing datasets of human-AI decisions with various AI assistance types. The state space incorporates human-centric and contextual factors, such as human skill, knowledge, and AI uncertainty. The action space comprises different types of human-AI interaction techniques. The reward function is crafted to capture objectives that are sparse or harder to capture, such as human learning or task enjoyment. Two experiments were conducted (N = 316 and N = 964) to compare the optimized policies against several baselines in AI-assisted decision-making. The study leverages offline RL to derive optimal support policies from existing datasets of human-AI decisions with various AI assistance.

Methodology

This track at the 15th Mediterranean Conference on Information Systems (MCIS) and The 6th MENA Conference on Information Systems (MENACIS) focuses on the intersection of Artificial Intelligence (AI) and Information Systems (IS), particularly through the lens of Design Science Research (DSR). The central methodological theme revolves around utilizing DSR to develop AI artifacts and treatments for design problems within IS. The track solicits research employing DSR to generate prescriptive knowledge derived from AI, aiming to empower IS across their entire lifecycle, from conception and design to development and operations. The research design emphasizes both artifact descriptions and evaluation results, encouraging submissions of completed research and work-in-progress papers with preliminary findings. Data collection methods are implicitly addressed through the call for evaluation protocols for AI-driven systems, suggesting a focus on empirical validation of AI artifacts and their impact on IS. Specific areas of interest include, but are not limited to, the development and evaluation of AI-based solutions for IS development, operation, quality assurance, security, usability, and software process management. Furthermore, the track encourages research on specific AI methodologies such as Business Analytics, Big Data, Machine Learning (including Deep Learning), Natural Language Processing, Computer Vision, and Knowledge Engineering, and their application within IS contexts.

Methodology

This research investigates the selection of human-centered design (HCD) methods by comparing selections made by novice design teams with recommendations generated by the large language model (LLM), GPT-3.5. The study aims to understand the differences between human and AI-driven method selection in design.

Research Questions:

- R1: Are methods selected by design teams similar or different to methods recommended by an LLM?
- R2: What similarities and differences exist between human- and LLM-sourced rationales for method selections?

Data Collection Methods:

The study employed a mixed-methods approach involving both quantitative and qualitative data collection and analysis. Data was collected from two primary sources:

1. **Novice Design Teams:** Method selections and justifications were sourced from 402 novice design teams participating in five offerings of a project-based engineering design course at a large public university. These teams were constrained to select from a curated set of 12-18 methods that varied across design phase.
2. **GPT-3.5:** The OpenAI GPT-3.5 API was used to generate design method recommendations in response to project- and design phase-specific prompts reflective of the novice teams' context. GPT-3.5 had no constraints on the methods available to it.

Data Analysis Methods:

1. **Summary Data Analysis of Method Selections:** The prevalence and differences in method selections between novice teams and GPT-3.5 were examined across different design phases and project types.
2. **Quantitative Data Analysis of Method Selection Justifications:** Latent Dirichlet Allocation (LDA) was used to identify prevalent topics within the justifications provided for method selections by both novice teams and GPT-3.5.

Key Findings:

The study observed that GPT-3.5 appears to represent design method knowledge held in method repositories well. However, GPT-3.5's method selection recommendations appeared to poorly distinguish between HCD phases and appeared limited to highly specific aspects of HCD phases.

Conclusions:

The findings highlight the unique contribution of human design cognition in design decision-making relative to LLMs, and herald the promise of human-AI teaming in design method selection.

Methodology

This research employs a participatory AI design approach, centered around the development and application of a web tool called "Deliberating with AI." The tool is designed to enable stakeholders to create and evaluate machine learning (ML) models using historical data. This process aims to facilitate the examination of strengths and shortcomings in past decision-making processes and to promote deliberation on improvements for future decisions. The design of the tool incorporates principles of participatory AI design, emphasizing stakeholder involvement in the creation of ML models. It also integrates elements of reflection and deliberation to encourage introspection and discussion during the model building and evaluation phases.

The study applies this tool within the context of graduate school admissions decisions, involving both faculty (decision-makers) and students (decision subjects) as stakeholders. The ML models generated through the tool serve as boundary objects, facilitating deliberation among stakeholders regarding organizational decision-making practices. The research methodology includes user studies to observe interactions and discourse among participants as they use the ML models to deliberate and envision fair decision-making processes. Data collected from these studies are analyzed to understand how the ML models support the sharing of perspectives and the discussion of nuances in admissions decision-making. The findings from this case study are intended to inform future research on stakeholder-centered participatory AI design and technology for organizational decision-making.

Methodology

Research Design and Data Collection Methods

This research critiques and builds upon the work of Anderson and Anderson (A&A) in machine ethics, specifically their prima facie duty approach, which computerizes W. D. Ross's (1930) ethical framework. A&A utilize inductive logic programming to discover decision principles within specific domains, such as healthcare scenarios. Their methodology involves associating potential actions (e.g., notifying or not notifying a supervisor) with ordered triples representing the degree to which duties like non-maleficence, beneficence, and autonomy are satisfied. Values are assigned to these duties (e.g., +1 for beneficence, -1 for its absence, +2 for respecting autonomy, -2 for violating it). The system then inductively seeks an equilibrium across various cases, guided by the moral intuitions of ethicists, to derive a coherent decision principle. This research identifies that A&A's system relies on the moral intuition of ethicists, using their judgments as training data to discover a coherent set of principles. The paper proposes an alternative non-intuition-based approach rooted in deontology, formulating ethical rules using quantified modal logic. The research aims to derive ethical principles from first principles based on formal properties of reasons for action, offering a more rigorous and transparent approach to machine ethics.

Methodology

This whitepaper employs a mixed-methods approach, combining legal analysis, ethical considerations, and policy recommendations to address the impact of Artificial Intelligence (AI) on human rights. The research design centers on the development of a comprehensive international framework, specifically the "Munich Convention on AI, Data, and Human Rights," to align AI governance with human rights principles.

Research Design:

- **Normative Analysis:** The research examines existing international legal instruments, particularly the UN Charter and the Universal Declaration of Human Rights, to establish a foundation for human rights-based AI governance.
- **Expert Consultation:** The "Munich Convention" was drafted with contributions from over 50 global experts, indicating a collaborative and iterative design process.
- **Policy Analysis:** The paper analyzes current AI regulatory frameworks, including the UN General Assembly’s Resolution on “Safe, Secure and Trustworthy AI,” the EU AI Act, and the Council of Europe’s ‘Framework Convention on AI, Human Rights, Democracy, and the Rule of Law,’ identifying their limitations in addressing the global digital divide and risks to marginalized communities.

Data Collection Methods:

- **Document Analysis:** The study relies on the analysis of international legal documents, resolutions, and conventions related to human rights and AI governance.
- **Expert Opinions:** The drafting of the "Munich Convention" involved gathering and synthesizing the opinions of international experts in AI ethics, law, and policy.
- **Case Studies:** The paper references concrete examples of AI deployments, such as robots at international borders and AI in gambling, to illustrate the potential for harm and the need for regulation.

Key Findings and Conclusions:

- Current AI regulatory frameworks are fragmented and insufficient to address the global digital divide and the disproportionate risks AI poses to marginalized communities.

- A unified and binding international framework, such as the proposed "Munich Convention on AI, Data, and Human Rights," is urgently needed to align AI governance with human rights principles.
- The UN Human Rights Council (UNHRC) is well-positioned to initiate global discussions on a binding convention for AI governance, grounded in human rights principles.
- The paper emphasizes the need to codify rights that empower individuals to opt out, be forgotten, seek explanations, and access remedies to maintain adequate human rights standards in the context of AI.

Participants and Sampling Techniques

Methodology

Participants and Sampling Techniques

This paper does not explicitly detail a study with human participants or specific sampling techniques. Instead, it presents a case study analysis of the Social Security Administration (SSA) and its implementation of Artificial Intelligence (AI) in the adjudication process. The "participants" in this context are the SSA as an organization, its adjudicators, and the claimants within the disability claims system. There was no sampling technique used. The study examines the entire system and its evolution.

Methodology

Participants and Sampling Techniques

This document presents a fictional case study and, therefore, does not involve actual participants or sampling techniques. It describes a hypothetical scenario in which the city of New Leviathan collaborates with the Wales Consulting Group (WCG) to develop a data-driven crime reduction program. The case study outlines the types of data that WCG would access, including public records, court filings, licenses, addresses, phone numbers, social media data, criminal databases, and police records. However, it does not detail any specific sampling methods or participant selection processes, as the focus is on the ethical considerations of using AI in public sector data analytics rather than empirical research. The "participants" in this scenario are effectively the residents of New Leviathan, whose data is being used, but they are not actively involved in a research study in the traditional sense.

Methodology

This paper employs a multiple case study methodology to analyze efforts aimed at operationalizing AI principles. Three cases were selected based on their scope, scale, and novelty to provide guidance for AI stakeholders. The cases are:

1. **Microsoft’s AI, Ethics and Effects in Engineering and Research (AETHER) Committee:** This case examines the development and function of the AETHER Committee in informing decisions about AI applications, including facial recognition. The committee comprises employees from across the company, organized into seven working groups.

2. **OpenAI’s Staged Release of GPT-2:** This case explores OpenAI's decision to release a powerful AI language model in stages over nine months in 2019, rather than all at once, to identify and address potential societal and policy implications.

3. **The OECD AI Policy Observatory:** This case investigates the OECD's initiative to facilitate international coordination in implementing AI principles through the AI Policy Observatory, launched in February 2020.

Data collection involved interviews, conversations, feedback, and review from individuals with expertise in AI governance. These individuals include Jared Brown, Miles Brundage, Peter Cihon, Allan Dafoe, Cyrus Hodes, Eric Horvitz, Will Hunt, Daniel Kluttz, Yolanda Lannquist, Herb Lin, Nick Merrill, Nicolas Miailhe, Adam Murray, Brandie Nonnecke, Karine Perset, Toby Shevlane, and Irene Solaiman.

Methodology

Participants and Sampling Techniques

This research involved human participants in the data collection phase to establish a baseline for human-human collaboration in decision-oriented dialogues. Specifically, dialogues were collected for three tasks: Assignment (conference reviewer matching), Planning (travel itinerary planning), and Mediation (group scheduling). The number of participants and specific sampling techniques used to recruit them are not explicitly detailed in this document. However, the collected human-human dialogues served as a reference point for evaluating the performance of language models. The average length of these dialogues was 13 messages over 8 minutes (Table 1, not included in the provided text).

Methodology

Participants and Sampling Techniques

This research employed a qualitative approach, gathering data through semi-structured interviews. A total of 47 participants were interviewed to investigate the influence of AI-sourced positive emotions on human teammates within human-AI teams (HATs). The sampling strategy and specific demographic information of the participants were not explicitly detailed within the provided text.

Methodology

Participants and Sampling Techniques

This study involved a sample of 199 participants recruited through Amazon Mechanical Turk. The experiment compared three cognitive forcing designs, two explainable AI approaches, and a no-AI baseline. The study also considered the participants' Need for Cognition (NFC), defined as an individual's motivation to engage in effortful mental activities, as a moderating factor in the effectiveness of the interventions.

Methodology

Participants and Sampling Techniques

This research employed human participants to evaluate the impact of automatically generated explanations on user perceptions of an Explainable AI (XAI) system. A total of 60 participants were recruited via Turk Prime to provide over 2000 samples of action-explanation pairs. The participants were involved in a data collection pipeline designed to gather "think-aloud" explanations as they interacted with a modified version of the Frogger game. This pipeline consisted of three phases: a guided tutorial to familiarize users with the interface, rationale collection where participants explained their actions during gameplay, and a review phase where participants could review and edit the action-explanation pairs. The use of Turk Prime suggests a convenience sampling approach, aiming for a diverse pool of participants readily available through the platform.

Methodology

Participants and Sampling Techniques

This research comprised five studies with a total of 731 participants (N=731). Specific sample sizes for each study were as follows: Study 1 (N=340), Study 2 (N=340), Study 3 (N=286), and Study 4 (N=114). The sampling techniques were not explicitly detailed in the provided text. Participants collaborated with a simulated AI in a maze-solving task across all studies.

Methodology

Participants and Sampling Techniques

This research involved two experiments with distinct participant groups. The first experiment included a sample of 316 participants (N = 316), while the second experiment comprised 964 participants (N = 964). The document does not specify the sampling techniques employed to recruit these participants.

Methodology

Participants and Sampling Techniques

This research employed human participants to collect data for training a neural network model designed to generate explanations for AI actions. A total of 60 participants were recruited via Turk Prime to provide over 2000 samples of action-explanation pairs. Participants engaged in a modified version of the Frogger game, a sequential decision-making task, where they simultaneously played the game and verbalized explanations for each action taken. This "think-aloud" protocol was implemented remotely.

Data Analysis Techniques

Methodology

Data Analysis Techniques

This research employs offline reinforcement learning (RL) to model human-AI decision-making and optimize human-AI interaction for diverse objectives. The problem of human-AI decision-making is cast as a Markov Decision Process, and policies are learned to optimize different objectives. Offline RL is chosen for its ability to model objectives that are sparse or harder to capture (e.g., human learning), capture human-centric and contextual factors (e.g., human skill, AI uncertainty) as part of the state space, and adapt the support with an action space comprised of different types of human-AI interaction techniques effective for specific contexts. Optimal support policies are derived from existing datasets of human-AI decisions with various AI assistance. The reward function and the construction of the state and action spaces are customizable to optimize various objectives in AI-assisted decision-making. The state space includes relevant characteristics, skill on the task, knowledge of the concept in question, and AI uncertainty. The action space consists of different types of AI assistance or intervention (e.g., no assistance, AI recommendation, AI explanation, on-demand assistance). The study focuses on optimizing decision accuracy and human learning about the task.

Methodology: Data Analysis Techniques

This workshop focuses on analytic modeling, specifically regression and optimization analysis, to develop skills in the analytic process. The curriculum includes data cleaning and mining, emphasizing the importance of clean, consistent, and reliable data, along with documenting the process and its impact. Statistical inference, including standard deviation and standard error, is also covered. Advanced methods such as machine learning and artificial intelligence are explored. Participants are encouraged to bring their own data to assess its use in improving decision-making processes. The course aims to provide participants with the ability to use data more effectively and apply modeling techniques to business problems.

Methodology: Data Analysis Techniques

This paper addresses the challenge of decision-making under uncertainty when the underlying data distribution is unknown. Traditional two-step procedures, involving machine learning for uncertainty modeling followed by stochastic optimization, are criticized for potential sub-optimality due to loss function mismatches and inconsistent assumptions between steps. The paper advocates for a direct data-driven approach, integrating these steps to optimize decisions based directly on historical data.

The proposed methodology adapts the framework of statistical decision theory. Key modifications include: (1) making no assumptions about the form of the underlying distribution, effectively framing the problem as nonparametric estimation; and (2) defining the loss function as the true economic cost associated with a decision, dependent on the random realization of the uncertain variable.

Instead of pursuing traditional optimality criteria that require distributional assumptions, the paper proposes minimizing the empirical risk, calculated as the average cost over the observed data, within a pre-determined hypothesis class of decision functions. This approach aims to generate efficient algorithms, particularly when the cost function is convex and the hypothesis class is linear. The paper suggests that even in more complex, non-convex situations, theoretical performance guarantees can be derived, bounding the sub-optimality with respect to the true risk and demonstrating convergence as the sample size increases.

The paper then applies this general methodology to the specific problem of electric power dispatch for renewable integration, where the uncertainty stems from the intermittent nature of renewable energy sources like wind power. The empirical risk minimization paradigm is used to develop algorithms for this application, and theoretical guarantees are derived, followed by empirical testing using data from the Bonneville Power Administration (BPA).

Methodology: Data Analysis Techniques

This research explores various methodologies for training decision-making artificial intelligence (AI) agents, focusing on techniques that leverage human guidance to enhance learning efficiency and safety. The study identifies and reviews five primary types of human guidance: evaluation, preference, goals, attention, and demonstrations without action labels. These methods are presented as complementary to, rather than replacements for, traditional reinforcement learning (RL) and imitation learning (IL).

Evaluative Feedback: This technique involves human trainers providing instantaneous positive or negative feedback on the agent's actions. The interpretation of this feedback within RL frameworks varies, with researchers considering it as additional reward signals, policy gradients, value functions, or advantage functions. Algorithms such as Advise, TAMER, and COACH have been developed based on these interpretations. The key advantage of evaluative feedback is its immediacy and frequency, particularly beneficial in scenarios with delayed and sparse rewards, such as games like chess or Go.

Preference Learning: In this approach, the AI agent presents pairs of learned behavior trajectories to a human trainer, who then indicates the preferred trajectory. This method is particularly useful when evaluation is only feasible at the end of a trajectory. Preference learning is often formalized as an inverse reinforcement learning problem, aiming to infer the unobserved reward function from human preferences. A significant research area within this technique is preference elicitation, which focuses on selecting trajectories that maximize the information gained from human input.

High-Level Goal Specification: This method involves human trainers specifying high-level goals for the agent to achieve. The agent then attempts to accomplish each goal through a sequence of low-level actions. This approach aligns with behavioral psychology studies indicating that humans primarily imitate high-level program structures. A promising combination involves using imitation learning for goal selection at a high level and RL for learning low-level actions to achieve those goals. The suitability of this approach depends on the relative effort required to specify goals versus providing demonstrations, as well as safety considerations.

Attention Guidance: This technique addresses the challenge of information selection in complex environments. As AI agents transition from simple digital environments to the real world, they face the same challenge as humans: selecting important information from a complex and potentially irrelevant feature set.

Data Analysis Techniques

The curriculum in Artificial Intelligence and Decision Making incorporates several subjects that focus on data analysis techniques. Specifically, **6.3720 Introduction to Statistical Data Analysis** is a core subject within the "Data-centric" area. Furthermore, **6.3900 Introduction to Machine Learning** is also categorized as data-centric, suggesting a focus on machine learning methodologies for data analysis. The course **6.8611 Quantitative Methods for Natural Language Processing** is listed as both an AI+D Advanced Undergraduate Subject and a Social and Ethical Responsibilities of Computing (SERC) subject, indicating the use of quantitative data analysis techniques within the domain of natural language processing. Finally, **6.C01 Modeling with Machine Learning: from Algorithms to Applications** is a SERC subject that focuses on modeling using machine learning algorithms.

Methodology: Data Analysis Techniques

This paper presents a survey and framework for the design and evaluation of Explainable AI (XAI) systems, drawing upon research from machine learning, visualization, and human-computer interaction (HCI). The core methodology involves a thorough review of XAI-related literature across these disciplines to categorize interpretable machine learning design goals and evaluation methods. The analysis identifies a mapping between design goals for different XAI user groups and their corresponding evaluation methods. Based on these findings, the authors develop a framework that provides step-by-step design guidelines paired with appropriate evaluation methods, aiming to facilitate iterative design and evaluation cycles within multidisciplinary teams. The paper also includes summarized tables of evaluation methods tailored to different goals in XAI research and recommendations for XAI design and evaluation derived from the literature review. The ultimate goal is to share knowledge and experiences of XAI design and evaluation methods across multiple disciplines, addressing the challenges posed by differing objectives and independent topics within XAI research.

Methodology: Data Analysis Techniques

This research employs a novel approach to data analysis by utilizing machine learning (ML) models not for direct decision-making automation, but as boundary objects to facilitate stakeholder deliberation and improve human decision-making processes. The core methodology involves the creation and evaluation of ML models using historical data related to organizational decisions. Specifically, the study uses a web tool called "Deliberating with AI" to enable stakeholders (decision-makers and decision-subjects) to build and assess these models. The analysis focuses on how stakeholders interact with the ML models, the discussions they generate, and the insights they derive regarding past decision-making practices. The interactions and discourse from user studies are analyzed to understand how participants used the ML models to deliberate and envision fairer decision-making. The study does not aim to create an objectively unbiased system, but rather to provide a method for users to deliberate on improving future personal and organizational decision-making through participatory AI design.

Methodology: Data Analysis Techniques

This research article investigates the impact of Artificial Intelligence (AI) on decision-making processes within organizations. The study examines how AI transforms internal operations across various departments, including transport and consultation management, by predicting demands, adjusting supply levels, and optimizing results. AI software facilitates data-driven choices, improves customer targeting, enhances development, and assists in customizing data and reports for strategic decisions. The analysis encompasses AI mechanisms such as automation efficiency and complex management, including handling multilayered problems, succession planning incorporation, and business decision support for risk monitoring, assessment, compliance, and scam recognition. The research considers the application of these techniques across departments like Human Resources, Accounts & Finance, and Sales & Marketing, as well as industries such as Healthcare, Finance, Consulting Business, Transportation, and Food & Beverage. The study also acknowledges the role of AI in government authorities' defining and implementation processes in automation. The research also considers the challenges and concerns in AI decision-making, such as data quality and the "black box" issue, where inputs and operations are not easily explainable.

Methodology: Data Analysis Techniques

This research introduces the Data Quality Funnel Model as a novel data management approach designed to enhance the accuracy, reliability, and value of data used in AI systems, ultimately improving business decision-making and flexibility. The model emphasizes the critical role of machine learning and predictive analytics in enabling effective business strategy and growth, contingent upon the quality of input data.

The study highlights several data analysis techniques crucial for effective data management within AI systems. Data pre-processing or cleansing is identified as a critical initial step, involving error and inconsistency elimination, normalization for standardized comparison, integration of data from various sources, and data fusion to merge multiple sources into a coherent analysis. The concept of Data-as-a-Service (DaaS) is presented, requiring continuous Data Quality Management processes utilizing machine learning models for quality assurance. The use of synthetic data, generated via mathematical models or algorithms, is also explored as a means to imitate real data for privacy, ethics, and overall data quality purposes, supporting applications such as machine learning training and internal business uses.

Furthermore, the research emphasizes the importance of Explainable AI (XAI) methods, such as LIME and SHAP, to enhance trust in automated decisions by providing explanations for AI recommendations. The study also touches on the use of machine learning and AI to enhance resource management in cloud computing.

The core finding is that prioritizing data quality through techniques like data cleansing, normalization, integration, and the adoption of models like the Data Quality Funnel, is essential for accountable AI systems and improved business outcomes. The research concludes that good data management and responsible AI reinforce each other, leading to more effective and competitive business strategies.

Methodology: Data Analysis Techniques

This research presents a visual analytics framework designed for analyzing data collected during the evaluation of interactive AI systems in clinical decision support. The framework aims to identify subgroups of users and patients based on their characteristics, detect outliers, and ensure adherence to regulatory guidelines. The system design was driven by early-stage clinical AI regulatory guidelines, specifically DECIDE-AI.

The framework integrates methodologies from multiple fields, including multiple-factor analysis and hierarchical clustering, to analyze and enhance AI effectiveness in clinical decision-making. Interactive visualizations are used to explore and interpret the results of these analyses. The framework is designed to be scalable and cost-effective, addressing the gap in advanced open-source options for specialized fields like clinical AI evaluation.

The effectiveness of the framework is demonstrated through a case study evaluating a prototype AI-based clinical decision-support system for diagnosing pediatric brain tumors. The framework emphasizes an exploratory and holistic approach to interpreting the results of AI evaluation studies, providing an overview of the collected data complemented by secondary views.

Results

Positive Impacts of AI on Decision Making

Results: Positive Impacts of AI on Decision Making

AI systems are increasingly deployed to assist human experts in decision-making across various high-stakes domains, including healthcare, criminal justice, and finance. Research indicates that even imperfect AI recommendations can lead to enhanced performance in human-AI teams compared to either the human or the AI operating independently (Wang et al., 2016; Jaderberg et al., 2019). This improvement stems from the potential for complementary expertise, where the human and AI can leverage each other's strengths and mitigate weaknesses.

The effectiveness of human-AI collaboration hinges on the human's ability to develop an accurate mental model of the AI system's capabilities, particularly its error boundary. This mental model allows the human to predict when the AI is likely to err and, consequently, decide when to accept or override the AI's recommendation. Studies using platforms like CAJA (Bansal et al., 2019) have demonstrated that properties such as parsimony and non-stochasticity of AI error boundaries significantly improve the human's ability to form accurate mental models. Parsimonious error boundaries, which are simple to represent and understand, and non-stochastic error boundaries, which exhibit predictable and reliable patterns of success and failure, facilitate the development of effective mental models.

Conversely, a mismatch between the human's mental model and the true error boundary can lead to suboptimal decisions, such as trusting the AI when it makes an erroneous recommendation or distrusting it when it is correct. These errors can negatively impact productivity and accuracy. Therefore, fostering accurate mental models is crucial for maximizing the benefits of AI-advised human decision-making.

Results: Positive Impacts of AI on Decision Making

AI-assisted decision making combines the strengths of humans and AI to produce superior outcomes compared to either acting alone (Zhang et al., 2020). AI can aid in decisions ranging from everyday choices to critical judgments in areas like healthcare, finance, and criminal justice. While full automation is often undesirable, particularly in high-stakes scenarios, the integration of AI offers significant potential for improved decision-making. Effective human-AI partnerships require users to understand AI capabilities to appropriately calibrate their trust, leading to improved decision outcomes (Bansal et al., 2019). Trust calibration, defined as the alignment between a person's trust in AI and its actual capabilities (Lee & Moray, 1994), is crucial for successful AI-assisted decision making. When trust is appropriately calibrated, individuals can leverage AI's strengths while mitigating its weaknesses, leading to more informed and effective decisions.

Positive Impacts of AI on Decision Making: Reducing Overreliance through Explanations

Research indicates that explanations provided by AI systems can positively impact decision-making by reducing overreliance, a phenomenon where individuals accept incorrect AI predictions without verification. This contradicts previous findings suggesting explanations either exacerbate or do not alter overreliance unless cognitive forcing functions are implemented. Vasconcelos et al. (2023) posit that individuals strategically choose whether to engage with AI explanations based on a cost-benefit analysis, weighing the cognitive effort required to complete a task independently or verify the AI's output against the perceived benefits.

The study introduces a theoretical framework suggesting that explanations reduce overreliance when tasks are cognitively demanding and explanations sufficiently reduce the cost of verification, or when explanations are easier to understand. This framework was tested through five experiments (N=731) using a maze-solving task with a simulated AI.

Study 1 (N=340) manipulated task difficulty, demonstrating that while explanations did not reduce overreliance in easy tasks, they did so as task difficulty increased. Studies 2 (N=340) and 3 (N=286) focused on the cognitive cost of understanding AI explanations, finding that lower cognitive effort led to a greater reduction in overreliance. These findings suggest that the effectiveness of AI explanations in mitigating overreliance is contingent on the relative cognitive costs associated with task completion and explanation comprehension. The methodology employed a simulated AI in a controlled experimental setting, allowing for precise manipulation of task difficulty and explanation complexity. The conclusion drawn is that carefully designed explanations, particularly in cognitively demanding tasks, can enhance human-AI collaboration by promoting more judicious reliance on AI systems.

Results: Positive Impacts of AI on Decision Making

This research investigates the accuracy-time tradeoffs of AI assistance in decision-making, particularly under time pressure. The central premise is that AI assistance can positively impact decision-making by both increasing accuracy and reducing decision-making time, especially in time-sensitive scenarios.

Key Findings:

- **Differential Impact of AI Assistance Types:** The study reveals that different types of AI assistance exhibit varying accuracy-time tradeoffs. For instance, AI assistance requiring more cognitive effort from the user tends to increase decision-making time but reduces overreliance on AI recommendations. Conversely, AI assistance requiring less cognitive effort, such as direct recommendations, can expedite response time but potentially lower accuracy due to increased overreliance.
- **Influence of Time Pressure:** Time pressure significantly affects how users interact with different AI assistance types. The research suggests that certain AI assistance modalities become more beneficial under time pressure compared to settings without time constraints. This implies that the effectiveness of AI assistance is context-dependent, necessitating adaptive strategies.
- **Predictive Power of Overreliance Rate:** A user's overreliance rate on AI recommendations emerges as a key predictor of their behavior. The study identifies distinct patterns in how individuals classified as "overreliers" and "not-overreliers" utilize different AI assistance types. This finding underscores the importance of considering individual differences in trust and reliance on AI when designing and deploying AI-assisted decision-making systems.

Methodology:

- **Experimental Design:** The research employed a controlled experimental design where participants completed logic puzzles under varying conditions of AI assistance and time pressure. Participants were assigned to different AI assistance conditions: no AI assistance, AI recommendation and explanation before decision-making ("AI-before"), AI recommendation and explanation after an initial decision ("AI-after"), or a mixed condition.
- **Time Pressure Manipulation:** Time pressure was induced through the use of on-screen timers, creating a sense of urgency for participants to complete the tasks within a specified timeframe.
- **Data Collection and Analysis:** The study collected data on accuracy, response time, and overreliance rate. Statistical analyses were conducted to examine the effects of AI assistance type and time pressure on these outcome measures.

Conclusions:

The study concludes that AI assistances exhibit different accuracy-time tradeoffs under time pressure compared to conditions without time pressure. The findings highlight the importance of adapting AI assistance strategies based on factors such as the type of assistance, the presence of time pressure, and individual user characteristics, particularly their tendency to over- or under-rely on AI recommendations. The research contributes to a deeper understanding of human-AI interaction in time-sensitive decision-making contexts and provides insights for designing more effective and adaptive AI-assisted systems.

Results: Positive Impacts of AI on Decision Making

This study investigates the evolution of human confidence in AI and in themselves, and their impact on human-AI decision-making, specifically in AI-assisted decision-making contexts where humans are responsible for the final decision. The research employs a cognitive study and a quantitative model to explore how changes in AI performance and resulting feedback affect human confidence in AI and self-confidence, and how these confidences are associated with the probability of accepting AI suggestions. The study uses a chess puzzle task, where participants collaborate with an AI named Taylor (based on the Stockfish chess engine), to make the best chess move given a board state.

The findings indicate that human self-confidence, rather than confidence in AI, primarily directs the decision to accept or reject AI suggestions. The research also reveals a tendency for humans to misattribute blame to themselves, potentially leading to a cycle of over-reliance on poorly performing AI systems. These insights highlight the need for effective calibration of human self-confidence to enhance successful AI-assisted decision-making.

Results: Positive Impacts of AI on Decision Making

This research investigates the accuracy-time tradeoffs of AI assistance in decision-making, particularly under time pressure. The central premise is that AI can positively impact decision-making by both increasing accuracy and reducing decision-making time, especially in time-sensitive scenarios.

Key Findings:

- **Differential Impact of AI Assistance Types:** The study reveals that different types of AI assistance exhibit varying accuracy-time tradeoffs. For instance, AI assistance requiring more cognitive effort from the user tends to increase decision-making time but reduces overreliance on AI recommendations. Conversely, AI assistance that requires less cognitive effort, such as direct recommendations, can expedite response time but may lead to increased overreliance, potentially compromising accuracy.
- **Influence of Time Pressure:** Time pressure significantly affects how users interact with AI assistance. The research indicates that certain AI assistance types become more beneficial under time pressure compared to settings without time constraints. Specifically, when AI is likely to be correct, relying on AI assistance can save time, proving advantageous in time-pressured situations.
- **Predictive Power of Overreliance Rate:** The study identifies a user's overreliance rate as a key predictor of their behavior when using AI assistance. Individuals prone to overreliance and those who are not utilize different AI assistance types differently, suggesting that personalized AI assistance strategies could be more effective.
- **Scarcity Effect:** The research explores the "scarcity effect," where participants rely more on AI assistance when it is presented less frequently. This finding suggests that strategically withholding AI assistance can influence user reliance and potentially optimize decision-making outcomes.

Methodology:

- **Experimental Design:** The research employed two experiments involving participants completing logic puzzles under varying conditions of AI assistance and time pressure.
- **AI Assistance Types:** Participants were exposed to different AI assistance types: (i) no AI assistance, (ii) AI recommendation and explanation before the user's decision ("AI-before"),

(iii) AI recommendation and explanation after the user's initial decision ("AI-after"), and (iv) a mixed condition with random variations of the three.

- **Time Pressure Manipulation:** Time pressure was induced through on-screen timers, both for the overall task and for individual puzzles, creating a sense of urgency.
- **Data Collection and Analysis:** The study measured accuracy, response time, and overreliance rate to assess the impact of AI assistance and time pressure on decision-making behavior.

Conclusions:

The research concludes that AI assistance has different accuracy-time tradeoffs under time pressure compared to no time pressure. The findings underscore the importance of adapting AI assistance strategies based on factors such as the user's overreliance tendencies and the specific demands of the task. The study suggests that personalized and context-aware AI assistance can optimize decision-making outcomes in time-sensitive environments.

Results: Positive Impacts of AI on Decision Making

This research investigates the conditions under which explanations provided by AI systems can positively impact human decision-making by reducing overreliance. Overreliance, defined as the tendency to accept AI predictions even when incorrect, is a prevalent issue in human-AI collaboration. Contrary to previous findings suggesting that explanations either exacerbate or do not alter overreliance without intervention, this study demonstrates that explanations can indeed reduce overreliance under specific circumstances.

The core argument is formalized within a cost-benefit framework, positing that individuals strategically evaluate the cognitive effort associated with completing a task independently versus verifying the AI's explanation, relative to the perceived risk of overreliance. The study hypothesizes that when tasks are cognitively demanding and explanations sufficiently reduce the cost of verification, or when explanations are inherently easier to understand, overreliance will decrease.

This framework was empirically tested through five studies (N=731) utilizing a maze-solving task where participants collaborated with a simulated AI. Study 1 (N=340) manipulated task difficulty, replicating prior findings of equivalent overreliance levels between AI predictions alone and AI predictions with explanations for easy tasks. However, as task difficulty increased, overreliance was significantly lower when explanations were provided. Studies 2 (N=340) and 3 (N=286) focused on manipulating the cognitive cost of understanding AI explanations, revealing that lower cognitive effort associated with explanations correlated with a greater reduction in overreliance. Study 4 (N=114) explored the balance between cognitive effort (costs) and incentives (benefits) in decision-making. Study 5 adapted the Cognitive Effort Discounting paradigm to quantify the utility of different explanations, providing further support for the cost-benefit framework.

The collective findings suggest that the previously observed null effects of explanations on overreliance may stem from a narrow focus on scenarios where the cognitive costs of the task and explanation are comparable and non-trivial. By manipulating these costs, the research demonstrates that explanations can be effective in reducing overreliance when they offer a relatively lower-cost alternative to independent task completion or blind reliance on the AI.

Positive Impacts of AI on Decision Making

Artificial Intelligence (AI) has emerged as a transformative force in organizational decision-making, enhancing proficiency and accuracy across various sectors. AI-driven approaches, leveraging machine learning and natural language processing, expedite business activities, leading to faster results and increased responsiveness (Favour Oluwadamilare Usman et al, 2024). This modernization simplifies processes, customizes operations, and ultimately increases client and customer satisfaction.

The integration of AI facilitates data-driven choices, improves customer targeting, and enhances development by customizing data and reports for strategic decisions. AI mechanisms, such as automation, complex management, and succession planning, assist in risk monitoring, assessment, compliance, and fraud recognition across departments like Human Resources, Accounts & Finance, and Sales & Marketing.

AI's ability to analyze vast datasets rapidly accelerates the examination of data, providing insights for further investigation and implementation. This capability is particularly valuable in dynamic business environments requiring quick decisions. AI systems can process large quantities of data and generate outputs in fractions of a second, saving time and effort, thereby increasing productivity (Muhammad Eid BALBAA, 2024). Furthermore, AI reduces manual labor and paper usage, enabling business leaders to allocate their time to critical thinking.

AI also enhances the accuracy and stability of data analysis, minimizing the potential for human bias. It avoids task duplication, enabling more efficient resource allocation. The technology facilitates accurate and rapid forecasting, reducing the need for human intervention (Muhammad Eid BALBAA, 2024). Behavioral analysis, particularly in assessing product or service satisfaction, is improved through AI-driven survey analysis, enabling timely corrective actions (Nauri Hicham, 2023).

AI's benefits extend across diverse industries, including consulting, healthcare, and logistics, as well as within departments such as sales & marketing, finance, and human resources. It also improves government regulations and procedures by making processes faster, error-free, and more accurate than human-led interventions.

Positive Impacts of AI on Decision Making: Cognitive Forcing Functions

While the integration of AI into decision-making processes aims to enhance performance, research indicates a prevalent issue of overreliance on AI suggestions, even when incorrect. This phenomenon undermines the potential benefits of human-AI collaboration, resulting in suboptimal outcomes compared to scenarios where individuals make decisions independently. Explainable AI (XAI), designed to mitigate overreliance by providing rationales for AI recommendations, has not consistently demonstrated substantial success in reducing this bias.

Buçinca, Malaya, and Gajos (2021) investigated the application of cognitive forcing functions as a means to promote more analytical engagement with AI-generated explanations. Drawing upon dual-process theory, which posits that individuals often rely on heuristic-based System 1 thinking, the study hypothesized that interventions prompting analytical System 2 thinking could reduce overreliance.

The researchers conducted an experiment (N=199) comparing three cognitive forcing designs with two simple XAI approaches and a no-AI baseline. The findings revealed that cognitive forcing functions significantly reduced overreliance compared to the simple XAI approaches. This suggests that actively prompting users to engage more deeply with AI explanations can lead to improved decision-making by reducing the tendency to blindly accept AI suggestions. However, the study also identified a trade-off, with participants assigning less favorable subjective ratings to the designs that most effectively reduced overreliance. Furthermore, the benefits of cognitive forcing interventions were found to be moderated by an individual's Need for Cognition (NFC), with those higher in NFC benefiting more.

In conclusion, the research demonstrates the potential of cognitive forcing functions to positively impact decision-making in AI-assisted contexts by mitigating overreliance. However, the study also highlights the importance of considering user experience and individual cognitive differences when designing such interventions to ensure both effectiveness and acceptability.

Results: Positive Impacts of AI on Decision Making

This research investigates the potential of offline reinforcement learning (RL) to optimize human-AI interaction for diverse objectives beyond mere decision accuracy. The study posits that current AI decision-support tools often provide fixed support, neglecting human-centric objectives such as skill improvement and collaboration. The central argument is that AI support should be dynamic, adapting to context and individual needs to optimize both decision accuracy and other specified human-centric objectives.

The methodology involves casting human-AI decision-making as a Markov Decision Process and learning policies that optimize different objectives using offline RL. This approach allows for modeling objectives that are sparse or harder to capture, such as human learning, as part of the reward function. Human-centric and contextual factors, including human skill, cognitive load, motivation, and AI uncertainty, are incorporated into the state space. The action space comprises different types of human-AI interaction techniques tailored for specific contexts.

The study instantiated this approach with two objectives: human-AI accuracy on the decision-making task and human skill improvement. Optimal support policies were derived from existing datasets of human-AI decisions with various AI assistance types. These policies were then compared against several baselines in AI-assisted decision-making across two experiments (N = 316 and N = 964).

The results consistently demonstrated that participants interacting with policies optimized for accuracy achieved significantly better accuracy and human-AI complementarity compared to those interacting with other types of AI support. This suggests that RL-optimized policies can effectively enhance decision-making accuracy in human-AI collaborations. While optimizing for human learning proved more challenging, the study indicates the potential for RL to improve human skills through adaptive AI support.

The research concludes that offline RL is a promising approach for modeling the dynamics of human-AI decision-making, leading to policies that can optimize human-centric objectives and provide novel insights into the AI-assisted decision-making space. Furthermore, the study emphasizes the importance of considering human-centric objectives beyond decision accuracy in AI-assisted decision-making, opening up the research challenge of optimizing human-AI interaction for such objectives. The findings suggest that dynamic AI support, tailored to individual needs and context, can lead to improved decision accuracy and potentially enhance human skills in collaborative decision-making environments.

Negative Impacts of AI on Decision Making

Negative Impacts of AI on Decision Making

AI decision-making in government, while offering benefits such as reduced costs and consistency, presents significant risks, particularly for vulnerable populations. These risks include inherent or developed biases within AI systems, leading to potential harm for individuals less equipped to navigate advanced technologies. The automation of high-volume decision-making can adversely affect recipients of government social programs who are least able to address errors.

The "Robodebt" scheme in Australia exemplifies these negative impacts. An AI system erroneously identified overpayments and calculated debts for over 450,000 social security beneficiaries due to methodological errors. This resulted in incorrect or inflated debt calculations, leading to severe repercussions for vulnerable, low-socioeconomic debtors, including loss of housing and food, mental health issues, and even suicide. The Robodebt debacle highlights the importance of sound technology design in decision-making methodology and adherence to administrative law principles. Users of the agency's services were unaware of the use of AI decision-making and enforcement, mistakenly assuming human oversight in the attribution and calculation of overpayments.

This flawed implementation of AI has reduced public trust in AI-supported government decision-making and underscored the need for appropriate regulatory frameworks. AI technology challenges traditional legal norms, administrative law, and regulatory design, necessitating a reconsideration of how AI decision-making in government should be regulated.

Results: Negative Impacts of AI on Decision Making

This research identifies several complications introduced by AI that challenge the efficacy of negligence law in compensating for injuries resulting from AI-assisted decisions (Selbst, 2020). These complications highlight potential negative impacts of AI on decision-making processes and subsequent accountability.

Key Findings and Methodologies:

- **Unforeseeability of AI Errors:** The study argues that AI errors are often unpredictable and difficult to account for, disrupting traditional notions of foreseeability in negligence law. The author distinguishes between different types of AI, implying varying degrees of predictability and explainability in their error patterns (Selbst, 2020). This unpredictability stems from the complex algorithms and statistical models underlying AI systems, making it challenging for human users to anticipate potential failures.

- **Limitations on Human-Computer Interactions:** The research emphasizes the physical and cognitive limitations at the interface between humans and AI. These limitations can hinder a user's ability to effectively monitor, interpret, and override AI recommendations, potentially leading to errors or suboptimal decisions. The study suggests that reliance on AI can create a "deskilling" effect, where human users become overly dependent on the technology and lose their critical judgment (Selbst, 2020).

- **AI-Specific Software Vulnerabilities:** The introduction of AI into decision-making processes creates new software vulnerabilities that were not previously present. These vulnerabilities can be exploited by malicious actors, leading to compromised decisions and potential harm. The study implies that the complexity of AI systems makes them susceptible to novel forms of cyberattacks and manipulation, posing a significant challenge to data security and decision integrity (Selbst, 2020).

- **Unevenly Distributed Injuries:** The statistical nature of AI and its potential for bias raise distributional concerns. The research suggests that AI systems can perpetuate and amplify existing inequalities, leading to disproportionate harm to certain groups. The study highlights the risk of algorithmic bias, where AI models trained on biased data produce discriminatory outcomes, resulting in unfair or unjust decisions (Selbst, 2020).

Conclusions:

The author concludes that the unique characteristics of AI may impede the ability of negligence law to adapt and provide adequate compensation for injuries resulting from AI-assisted decisions. The study suggests that AI errors are becoming statistical certainties embedded within the technology, rather than individual operator faults. The author proposes that legislative interventions requiring interpretability and transparency in AI development, along with alternative regulatory approaches, may be necessary to prevent uncompensated losses and ensure accountability in the age of AI (Selbst, 2020).

Negative Impacts of AI on Decision Making

AI systems can perpetuate and amplify existing societal biases, leading to discriminatory outcomes in decision-making processes. Research indicates that biases embedded in training data, often reflecting historical inequalities, can result in AI models that unfairly disadvantage certain demographic groups. A 2019 study by the National Institute of Standards and Technology (NIST) revealed that commercial facial recognition applications exhibited significantly higher error rates in identifying individuals of Asian and African American origin compared to white men (Spring & Smith, n.d.).

The issue stems from multiple factors. First, training datasets may be skewed, predominantly featuring data from specific demographic groups, such as white males, leading to models that are less accurate and reliable for other populations (Spring & Smith, n.d.). Second, even when explicit demographic identifiers like race are removed from datasets, correlated features such as ZIP codes can still encode and perpetuate biases (Spring & Smith, n.d.). AI systems, designed to identify patterns, may inadvertently learn these correlations, resulting in discriminatory outcomes. For example, AI systems used to determine mortgage rates, trained on historical data reflecting past discriminatory lending practices against Latino and Black borrowers, are likely to perpetuate these biases, even without explicitly considering race (Spring & Smith, n.d.).

Furthermore, the lack of transparency in AI decision-making processes can make it difficult to detect and mitigate biases. As Smith (Spring & Smith, n.d.) suggests, removing demographic identifiers from data does not eliminate bias; it merely obscures the ability to track and address it. Therefore, rigorous testing and evaluation of AI systems are crucial to identify and mitigate potential biases before deployment. This includes thoroughly understanding the provenance of the data, its collection methods, and the potential consequences of its use. Datasheets for Datasets, as proposed in the literature, offer a structured approach to scrutinizing data and understanding its limitations (Spring & Smith, n.d.).

Negative Impacts of AI on Decision Making

Algorithmic bias in decision-making is defined as the use of algorithms that cause a systematic skew, resulting in unfair outcomes. This bias can manifest as discrimination under UK equality law when it leads to unfair treatment based on protected characteristics. However, algorithmic bias can also lead to unfair outcomes that are non-discriminatory. The core issue is that fairness encompasses more than just the absence of bias, and multiple, potentially incompatible, concepts of fairness exist.

Algorithms are often built on historical data to derive insights, prioritize resources, and assess risks, which can perpetuate existing biases. Examples include predictive mapping, individual risk assessment, and data scoring tools used in policing.

A significant challenge in addressing algorithmic bias is the lack of available data on protected characteristics needed to monitor outcomes and identify bias. Organizations often avoid collecting this data due to concerns about data protection law, customer reluctance to share data, and the potential for the data to reveal biased organizational outcomes.

The lack of transparency in the use of algorithms, particularly in the public sector, is a problem for trust. This lack of transparency hinders individual teams/projects from being transparent on their own and impedes democratic accountability for public sector decisions.

To address these issues, organizations are advised to build diverse, multidisciplinary internal capacity, understand the capabilities and limits of algorithmic tools, carefully consider whether individuals will be fairly treated by the decision-making process, make conscious decisions on appropriate levels of human involvement, put structures in place to gather data and monitor outcomes for fairness, and understand their legal obligations. Integrated impact assessments should be conducted before adopting new data analytics software.

Results: Negative Impacts of AI on Decision Making

AI systems, while offering potential efficiencies, present risks of unintentional bias that can negatively impact decision-making processes and organizational reputation. This risk is exacerbated by the limited explainability of many AI algorithms, making it difficult to identify and rectify embedded biases.

Key Findings and Methodologies:

- **Suboptimal Performance:** Algorithmic bias can lead to suboptimal application performance, resulting in missed opportunities and financial losses. For instance, biased lending practices can unfairly discriminate against certain populations, leading to both financial loss and missed financial gain.
- **Lack of Explainability:** The "black box" nature of many AI systems makes it challenging for managers to trust key business decisions based on AI inferences, especially when biases are embedded within training data in non-obvious ways.
- **Reinforcement of Bias:** If the output of an AI application influences its input, it can create a self-perpetuating cycle of bias, leading to continuously poor predictive performance.
- **Sources of Bias:** The paper identifies three general classes of bias: (1) those related to mapping business intent into the AI implementation, (2) those arising from the distribution of samples used for training (including historical effects), and (3) those present in individual input samples.
- **Proxy Goals and Feature Selection:** The use of proxy goals, where the AI objective is indirectly related to the actual business goal, can introduce bias. Similarly, the selection of input features can significantly influence decisions, with even seemingly innocuous features containing hidden biases. The exclusion of potentially favorable features can also lead to biased predictions.
- **Surrogate Data:** The reliance on surrogate data, such as using credit scores as a proxy for reliability, can cause information loss and bias the problem mapping. Surrogate data can also serve as proxies for restricted input, such as using zip codes as a proxy for race.
- **Data Set Issues:** Problems with training and production datasets can introduce bias. Unseen cases, where the system encounters data it was not trained on, can lead to "hidden"

mispredictions. Mismatched datasets, where production data differs significantly from training data, can result in poor model performance.

- **Manipulated Data:** Training data can be manipulated to skew results, as demonstrated by the Tay chatbot incident, where the system mimicked hate speech from its Twitter interactions.

Conclusions:

The paper highlights the significant negative impacts of bias in AI systems on decision-making processes. The lack of explainability, potential for reinforcing bias, and various sources of bias related to goal representation, feature selection, data sets, and data manipulation contribute to suboptimal performance and potential harm. The findings underscore the need for developing best practices to mitigate and manage bias in AI deployments to improve outcomes and foster confidence in AI results.

Negative Impacts of AI on Decision Making

The advent of artificial intelligence (AI) has amplified the impact of biased decision-making, extending its reach far beyond the localized effects of human bias. While human decision-making has historically been susceptible to bias, the deployment of AI algorithms on a global scale has resulted in adverse effects on larger groups of consumers and potential employees. Examples of such negative impacts include: disparate pricing for mortgages based on race, gender bias in corporate hiring, racial bias in prison sentencing risk assessments, and inaccuracies in facial recognition systems related to race and ethnicity.

Research has demonstrated the existence of bias in human decision-making, with studies showing that identical descriptions of male and female managers were evaluated differently based solely on gender. Similarly, resumes with Black-sounding names received significantly fewer interview requests compared to those with White-sounding names.

The core issue of algorithmic bias often stems from the data used to train AI models. Bias can arise from biased underlying data, such as historical judgements unfavorable to specific demographics, or from under-representation of certain demographics in the training dataset. For example, speech recognition systems trained on datasets with limited voice samples from Asian women may exhibit comprehension errors when processing their speech.

The risks associated with biased AI algorithms extend to regulatory and reputational domains. Companies have faced legal action and financial penalties for the unlawful use of biased AI systems, particularly in areas such as facial recognition technology. In response, some organizations are proactively limiting the application of such algorithms.

Results: Negative Impacts of AI on Decision Making

Artificial intelligence (AI) systems, particularly those employing machine learning, are increasingly utilized to automate judgments and make decisions affecting various aspects of life, including hiring processes, credit offers, targeted advertising, and allocation of government resources (Smith & Rustagi, 2020). While AI has the potential to reduce human subjectivity, it can also embed biases, leading to inaccurate or discriminatory predictions and outputs for specific population subsets (Smith & Rustagi, 2020).

One illustrative example detailed in the playbook involves a Bay Area tech firm that implemented an AI system to streamline its hiring process (Smith & Rustagi, 2020). The system, trained on data from current employees and past applicants, was intended to be "gender-blind" and "race-blind."

However, it consistently recommended predominantly white male candidates for interviews. Further investigation revealed that the AI system penalized resumes containing words associated with women, demonstrating a learned bias that the developers were unable to rectify. Consequently, the system was ultimately scrapped (Smith & Rustagi, 2020).

This case highlights how bias can infiltrate AI systems through data and throughout the development and evaluation stages of algorithms (Smith & Rustagi, 2020). The playbook emphasizes that bias in AI is not solely a technical issue but a broader business concern requiring oversight from business leaders to mitigate risks and lost opportunities (Smith & Rustagi, 2020). The authors argue that organizations currently lack the necessary frameworks to effectively address bias in AI, despite its increasing deployment in decisions affecting individuals and society (Smith & Rustagi, 2020).

Negative Impacts of AI on Decision Making

Recent research indicates that the integration of AI into decision-making processes, while intended to enhance performance, can paradoxically lead to suboptimal outcomes due to overreliance on AI suggestions (Buçinca, Malaya, & Gajos, 2021). Studies have shown that human+AI teams, despite generally outperforming individuals, often underperform compared to the AI acting autonomously, suggesting a failure to effectively combine human and artificial intelligence (Buçinca et al., 2021). Specifically, individuals tend to accept AI recommendations even when they are demonstrably incorrect, overriding their own potentially superior judgment (Buçinca et al., 2021).

Explainable AI (XAI), designed to mitigate this overreliance by providing rationales for AI decisions, has not proven substantially successful in reducing this effect (Buçinca et al., 2021). Even with explanations, individuals still make poorer decisions when the AI suggests incorrect solutions compared to scenarios without AI assistance (Buçinca et al., 2021). This phenomenon can be understood through the lens of dual-process theory, which posits that individuals primarily operate using System 1 thinking, relying on heuristics and cognitive shortcuts (Buçinca et al., 2021). The cognitive effort required to analytically evaluate each AI explanation is substantial, leading individuals to develop generalized heuristics about the AI's competence rather than engaging with the specific reasoning behind each recommendation (Buçinca et al., 2021). In essence, explanations may be interpreted as a general signal of competence, thereby increasing trust and, consequently, overreliance on the AI, irrespective of the explanation's actual validity (Buçinca et al., 2021).

Buçinca et al. (2021) explored the use of cognitive forcing functions—interventions designed to disrupt heuristic reasoning and promote analytical thinking—to counteract this overreliance. Their experimental results (N=199) demonstrated that cognitive forcing functions significantly reduced overreliance compared to simple XAI approaches (Buçinca et al., 2021). However, this reduction did not eliminate overreliance entirely, and a trade-off emerged between the effectiveness and acceptability of these interventions (Buçinca et al., 2021). Interventions that most effectively reduced overreliance were perceived as more difficult and less usable (Buçinca et al., 2021). Furthermore, the study revealed that individuals with a higher Need for Cognition (NFC) benefited more from cognitive forcing functions, suggesting that cognitive motivation moderates the effectiveness of XAI solutions and potentially exacerbates inequalities in AI-assisted decision-making (Buçinca et al., 2021).

Negative Impacts of AI on Decision Making: Overreliance

A significant challenge in human-AI collaboration is *overreliance*, where individuals accept incorrect AI decisions without verification, thereby diminishing their own agency and accountability

(Vasconcelos et al., 2023). This phenomenon is particularly concerning in high-stakes domains, potentially reinforcing machine bias under the guise of human agency (Vasconcelos et al., 2023). While Explainable AI (XAI) has been proposed as a mitigation strategy, empirical evidence has largely failed to demonstrate that explanations effectively reduce overreliance (Vasconcelos et al., 2023). Some research suggests that explanations may even exacerbate overreliance by increasing trust or anchoring humans to the AI's prediction (Vasconcelos et al., 2023). Interventions such as cognitive forcing functions, designed to compel deeper engagement with AI explanations, have shown some success in reducing overreliance, suggesting that unaided explanations are often insufficient (Vasconcelos et al., 2023).

Vasconcelos et al. (2023) challenge the notion that overreliance is an inevitable consequence of human cognition, arguing instead that individuals strategically choose whether to engage with AI explanations based on a cost-benefit analysis. Their research, comprising five studies (N=731), manipulates factors such as task difficulty, explanation difficulty, and monetary compensation within a maze-solving task to assess their impact on overreliance. The findings indicate that increased task difficulty (Study 1) and reduced explanation difficulty (Studies 2 and 3) can lead to a reduction in overreliance. Furthermore, Study 5, adapting the Cognitive Effort Discounting paradigm, quantifies the utility of different explanations, providing further support for the cost-benefit framework. The authors conclude that the ineffectiveness of explanations in reducing overreliance, as observed in previous studies, may be attributable to the explanations not sufficiently reducing the cognitive costs associated with verifying the AI's prediction (Vasconcelos et al., 2023).

Results: Negative Impacts of AI on Decision Making

Research indicates that despite the potential benefits of AI in human-AI teams, particularly in high-stakes domains, several factors can negatively impact decision-making processes. While AI accuracy is often the primary focus in development, it does not guarantee optimal team performance (Yin, Vaughan, and Wallach 2019; Lai and Tan 2018). A critical factor influencing team performance is the human's mental model of the AI system's capabilities, specifically its error boundary. A mismatch between the human’s mental model and the true error boundary can lead to suboptimal decisions, such as trusting the AI when it makes an erroneous recommendation or distrusting it when it makes a correct recommendation. These decision errors can lower productivity and/or accuracy.

The study identifies properties of an AI's error boundary, such as parsimony and non-stochasticity, that affect a human's ability to form an accurate mental model. Parsimonious error boundaries, which are simple to represent, and non-stochastic error boundaries, which can be modeled with a small set of features that reliably distinguish successes from errors, are crucial for effective human-AI collaboration. Task dimensionality, characterized by the number of features defining each instance, also influences the human's ability to create a mental model of the error boundary.

Controlled user studies, utilizing the C AJA platform, demonstrated that parsimony and non-stochasticity of error boundaries improve people's ability to create accurate mental models. The experiments also characterized how people create and update mental models over time, highlighting the need for guidance in this process. These findings underscore the importance of considering properties beyond accuracy to improve overall human-AI team performance.

Impact on Specific Decision-Making Contexts

Results: Impact on Specific Decision-Making Contexts

The application of Artificial Intelligence (AI) in health, healthcare, and biomedical science is rapidly expanding, influencing decision-making across various contexts. AI methods are increasingly employed in disease screening and detection (Ardila et al., 2019; Gulshan et al., 2016; Rajpurkar et al., 2022), prediction of real-time clinical outcomes, personalized treatment strategies, and enhancement of patient engagement. Furthermore, AI is being utilized to streamline administrative processes, reduce the burden of clinical documentation, and accelerate therapeutic drug discovery (Cai et al., 2015; Glover et al., 2022). Certain AI applications are designed to assist with human-like tasks, thereby providing input for human decision-making processes. However, the reliance on datasets and models influenced by human biases introduces the potential for implicit and explicit biases, particularly affecting under-represented groups (Obermeyer et al., 2019; Christensen et al., 2021). This can exacerbate existing inequities and create new disparities in healthcare outcomes. Additional risks associated with AI implementation include misdiagnosis, overuse of resources, privacy breaches, and workforce displacement or inattention due to over-reliance on AI (Matheny et al., 2022). The evolving nature of some AI models, which learn and adapt over time, can also lead to user confusion regarding their underlying logic. To address these challenges, the National Academy of Medicine is developing a draft framework for an AI Code of Conduct, comprising harmonized principles and simple rules, to guide responsible AI development and application across the AI lifecycle. This framework aims to equip stakeholders with the awareness and guidance necessary to make responsible decisions in real-time, promoting trustworthy AI in health, healthcare, and biomedical science. The framework is grounded in the core principles of a Learning Health System (LHS), emphasizing safety, effectiveness, patient-centeredness, timeliness, efficiency, equity, transparency, accountability, and security.

Impact on Specific Decision-Making Contexts

This document, "Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril," published by the National Academy of Medicine (NAM), addresses the potential impact of AI and machine learning on various decision-making contexts within healthcare. The publication aims to provide relevant information to physicians, nurses, clinicians, data scientists, healthcare administrators, public health officials, policymakers, regulators, purchasers, and patients. It covers key definitions, concepts, applicability, potential pitfalls, rate-limiting steps, and future trends related to AI in healthcare.

The document highlights the potential of AI to analyze and integrate massive amounts of data from diverse sources into clinical care. It also explores how AI innovations can facilitate population health models and interventions addressing social determinants of health. Furthermore, the publication discusses opportunities to advance precision medicine equitably and inclusively, reduce the cost of care delivery within healthcare organizations and businesses, and enhance interactions between healthcare professionals and patients, families, and caregivers. The role of legal statutes in influencing the adoption of AI in healthcare is also considered.

The publication emphasizes the importance of capturing data from patient encounters to create a "collective memory" of health services, potentially leading to a virtuous cycle of continuous improvement and innovation in healthcare delivery.

Results: Impact on Specific Decision-Making Contexts

This white paper, informed by a working group of six health systems and individual interviews, investigates the impact of AI governance on decision-making within healthcare systems. The research indicates that AI applications are already influencing faster triage and diagnosis, personalized treatment plans, streamlined clinical operations, patient communication, and resource allocation. However, the integration of AI tools raises ethical, legal, and practical concerns, necessitating comprehensive governance systems.

Key Findings:

- **Enhanced Decision-Making:** AI tools have demonstrated the potential to improve decision-making processes across various healthcare domains, including clinical decision support, population health management, operational optimization, and payment management.
- **Importance of Governance:** Effective AI governance systems are crucial for ensuring patient safety, maintaining ethical standards, ensuring regulatory compliance, fostering trust through transparency and accountability, and managing privacy concerns and other legal issues.
- **Organizational Alignment:** AI governance bodies require open communication with leadership and decision-makers to ensure that recommendations regarding AI tool assessment are taken seriously.
- **Multidisciplinary Approach:** Effective governance teams are multidisciplinary, ensuring a holistic approach to AI tool evaluation and implementation.
- **Governmental Influence:** Federal and state actions, such as the California Attorney General's request for a list of commercial decision-making tools and the National Institute of Standards and Technology's (NIST) AI Risk Management Framework, incentivize responsible AI governance. The Office of Civil Rights at HHS also released a draft rule in 2022 and finalized the rule in May 2024 regarding Section 1557 of the Affordable Care Act, "which prohibits discrimination … in covered health programs or activities." Part of this new rule focuses on discrimination resulting from the use of patient care decision support tools. Health systems must make a "reasonable effort" to identify and mitigate the risk of discrimination or inequitable care resulting from the use of these tools.
- **Liability Concerns:** FDA Commissioner Robert Califf emphasized the need for health systems to enhance AI governance, warning that they may face liability when algorithms produce incorrect results.

Methodology:

The research methodology involved:

- Convening a working group of six health systems with established AI governance systems.
- Conducting individual interviews with other health systems to understand their approaches to AI governance.

Conclusions:

The study concludes that AI governance is a relatively new but essential concept for health systems. While governance structures can vary, commonalities exist that facilitate effective governance processes. The findings underscore the need for health systems to prioritize the establishment of AI governance processes to ensure the safe, responsible, fair, and effective use of AI tools in healthcare.

Impact on Specific Decision-Making Contexts

This book chapter compilation addresses the impact of Explainable AI (XAI) across a diverse range of decision-making contexts. Specifically, the chapters delve into the application of XAI in healthcare (Chapter 8), finance (Chapter 9), legal and regulatory compliance (Chapter 11), autonomous systems (Chapter 12), customer service (Chapter 14), education (Chapter 15), social media (Chapter 16), environmental monitoring (Chapter 17), supply chain management (Chapter 18), and risk assessment and decision support (Chapter 19). The inclusion of these specific domains suggests a focus on how XAI can enhance transparency, accountability, and user understanding within each context.

The chapter on healthcare (Chapter 8) highlights the potential of XAI to improve patient understanding, implying a focus on how explanations can empower patients to make informed decisions about their treatment. Similarly, the chapter on finance (Chapter 9) emphasizes ensuring transparency and compliance, suggesting an exploration of how XAI can aid financial institutions in adhering to regulations and building trust with stakeholders. The chapter on autonomous systems (Chapter 12) focuses on ensuring safety and reliability, indicating an investigation into how XAI can provide insights into the decision-making processes of autonomous systems, thereby mitigating risks.

The inclusion of chapters on customer service (Chapter 14), education (Chapter 15), and social media (Chapter 16) suggests an exploration of how XAI can improve user experience, support learning, and enhance user trust and engagement, respectively. These chapters likely examine how explanations can be tailored to different user groups to facilitate better understanding and decision-making. Furthermore, the chapters on environmental monitoring (Chapter 17) and supply chain management (Chapter 18) suggest a focus on how XAI can provide insights for sustainability and enhance efficiency and visibility, respectively. Finally, the chapter on risk assessment and decision support (Chapter 19) directly addresses the application of XAI in improving the quality and reliability of decision-making processes.

The book, as a whole, aims to provide actionable insights into interpretable machine learning models, which are crucial for understanding the impact of XAI on these decision-making contexts. The emphasis on human-centric design (Chapter 6) suggests that the book will also explore how explanations can be tailored to the needs and preferences of different stakeholders, thereby maximizing the effectiveness of XAI in supporting decision-making.

Impact on Specific Decision-Making Contexts

AI-driven decision-making significantly impacts various organizational functions and industries. The integration of AI enhances the speed and efficiency of business activities through machine learning and natural language processing, enabling faster results and increased responsiveness in decision-making (Favour Oluwadamilare Usman et al, 2024). AI facilitates a more responsive and buyer-centric approach to modernization by exploring client responses and understanding preferences (Favour Oluwadamilare Usman et al, 2024).

AI implementation simplifies the decision-making process by processing large data quantities rapidly, saving time and effort, and increasing productivity (Muhammad Eid BALBAA, 2024). It reduces manual work, minimizes paper usage, and enhances data accuracy and stability. AI also prevents task duplication, allowing business leaders to focus on critical thinking and strategic decisions. Furthermore, AI enables accurate and fast advance forecast reports with reduced human intervention (Muhammad Eid BALBAA, 2024) and facilitates behavioral analysis for product or service satisfaction improvements (Nauri Hicham, 2023).

Specific industries benefiting from AI include consulting, healthcare, and logistics, while departments such as sales & marketing, finance, human resources, and procurement also experience improvements. AI is also valuable in government regulations and procedures, making processes faster, error-free, and more accurate than human intervention.

However, challenges exist, including the need for high-quality data to avoid adverse effects from inadequate, incorrect, or influenced data. Precise checking and authentication are necessary to overcome this challenge. Data clarity is crucial for explaining AI-driven results, requiring contributors to justify the outcomes using skeptical techniques (Muhammad Eid BALBAA, 2024). The "black box" issue in AI, where inputs and operations are opaque, also presents a challenge.

Results: Impact on Specific Decision-Making Contexts

This research investigates the impact of cognitive forcing functions on reducing overreliance on AI in AI-assisted decision-making contexts. The study, involving 199 participants, compared three cognitive forcing designs to two simple explainable AI (XAI) approaches and a no-AI baseline. The central finding is that cognitive forcing functions significantly reduced overreliance on AI compared to the simple XAI approaches. This suggests that interventions designed to compel analytical engagement with AI-generated explanations can be more effective than simply providing explanations.

However, the study also revealed a critical trade-off: the designs that most effectively reduced overreliance received the least favorable subjective ratings from participants. This indicates a potential usability and acceptability challenge associated with cognitive forcing functions, where increased cognitive effort may lead to decreased user satisfaction. Despite the reduction in overreliance, cognitive forcing functions did not eliminate it entirely, suggesting that participants still relied on incorrect AI predictions in instances where they would have made better decisions independently.

Furthermore, the research explored potential intervention-generated inequalities by examining the impact of cognitive forcing functions on individuals with varying levels of Need for Cognition (NFC). The results indicated that, on average, cognitive forcing interventions benefited participants with higher NFC more than those with lower NFC. This suggests that individual differences in cognitive motivation moderate the effectiveness of XAI solutions, and that cognitive forcing functions may exacerbate existing disparities in decision-making performance.

In summary, the study demonstrates the potential of cognitive forcing functions to mitigate overreliance on AI, but highlights the importance of considering usability, acceptability, and potential disparities in effectiveness across individuals with different cognitive profiles.

Impact on Specific Decision-Making Contexts

This course, Artificial Intelligence in Healthcare (BIODS 220), aims to equip students with the knowledge and skills necessary to understand and contribute to the rapidly evolving field of AI in healthcare. The curriculum is structured to provide a broad understanding of opportunities for AI in healthcare, fluency in cutting-edge deep learning algorithms, and practical abilities to develop models for diverse types of healthcare data. A key focus is placed on understanding real-world considerations and challenges associated with deploying AI algorithms in healthcare settings.

The course addresses the application of AI across various healthcare domains, including biomedical image interpretation (Wu et al. 2019; Liu et al. 2017), clinical event prediction (Harutyunyan et al. 2019), genomic analysis (Zhou et al. 2015; Poplin et al. 2016), drug discovery and drug interaction prediction (Torng et al. 2019; Ryu et al. 2018), intelligent healthcare spaces and environments (Yeung et al. 2019), and mobile health and wearables (Menictas et al. 2019; Tariq et al. 2018). Furthermore, the course explores recent applications of AI in addressing the COVID-19 crisis (Yan et al. 2020; Jumper et al. 2020; Harmon et al. 2020).

The course also emphasizes the importance of addressing challenges related to uncertainty and AI/human collaboration (Rosenberg et al. 2018), bias and fairness (Obermeyer et al. 2019), and privacy and security (Price et al. 2019) in the deployment of AI in healthcare.

The course structure is divided into two parts: developing deep learning algorithms for health data and deploying AI for health. The grading is based on assignments and a course project, which allows students to pursue an in-depth AI and healthcare project of their choice.

Results: Impact on Specific Decision-Making Contexts

AI's integration into higher education has evolved significantly, impacting various decision-making contexts. In the early 2000s, AI-integrated Learning Management Systems (LMSs) facilitated structured learning experiences by aiding educators in organizing course materials and assessments (Cavus, 2010). The 2010s witnessed a surge in personalized learning applications, with AI algorithms recommending tailored courses and content based on individual student needs (Maghsudi et al., 2021). Adaptive learning systems adjusted content difficulty based on student progress (Pratama et al., 2023). Mid-2010s saw the rise of AI-powered chatbots and virtual assistants, providing 24/7 support by answering student inquiries and guiding them through administrative processes (Chen et al., 2023; Essel et al., 2022; Lin, 2023). Late 2010s marked the adoption of predictive analytics to identify at-risk students, enabling targeted interventions based on data analysis of attendance, grades, and engagement (Seidel & Kutieleh, 2017; Shilbayeh & Abonamah, 2021). The 2020s introduced AI-powered proctoring tools for online exams (Kurni et al., 2023) and AI-generated content and automated grading systems (Ramesh & Sanampudi, 2022; Schroeder et al., 2022). These advancements influence decisions related to curriculum development, student support services, assessment strategies, and resource allocation within higher education institutions.

Results: Impact on Specific Decision-Making Contexts

The study by the ISB Institute of Data Science (IIDS) and Center for Business Markets (CBM) at the Indian School of Business investigates the landscape of Artificial Intelligence (AI) in the Indian healthcare sector. The research methodology combines primary data gathered through in-person interviews with leading hospitals across India and secondary research focusing on healthcare start-ups. The study identifies innovative AI technologies currently being explored in India to enhance healthcare quality and patient care. It also elucidates the challenges hindering the widespread adoption of AI-based technologies within the Indian healthcare system. The research findings suggest that AI has the potential to revolutionize healthcare delivery, particularly in addressing the scarcity of quality, affordable, and equitable healthcare services. Specific use cases and applications of AI in primary, secondary, and tertiary healthcare sectors are examined, highlighting the potential for AI to improve diagnostics, treatment, and overall patient outcomes. Furthermore, the study provides a roadmap for hospitals to adopt and integrate AI solutions, for start-ups to develop AI solutions addressing healthcare challenges, and for the government to promote AI adoption through electronic health records, capacity building, and regulatory frameworks. The research underscores the potential of AI to mitigate the economic burden of diseases, particularly non-communicable diseases and mental disorders, which are projected to cost India significantly. The study cites Qure.ai, a Mumbai-based start-up, as an example of a game changer in diagnosing Tuberculosis.

Results: Impact on Specific Decision-Making Contexts

AI decision-making in government contexts, while offering benefits such as cost reduction, consistency, and potential bias reduction, presents significant risks, particularly for vulnerable populations. A case study involving the "Robodebt" scheme in Australia revealed that flawed AI system design and methodology led to incorrect debt calculations for over 450,000 individuals, resulting in severe repercussions, including loss of housing and mental health issues. This instance highlights the critical importance of adhering to administrative law principles, such as legality, procedural fairness, and accountability, even when employing AI technologies. The lack of transparency regarding the use of AI in decision-making contributed to public distrust, as recipients assumed human oversight in debt attribution and calculation. The "Robodebt" debacle underscores the need for appropriate regulatory frameworks and systems for AI in government to mitigate harm to vulnerable populations and maintain public trust. The authors propose regulatory innovations, including a "Golden Rule" for negative decisions and a monitoring body to ensure fairness, accountability, and appropriate AI functioning, aligning with the "right to a well-calibrated decision."

Discussion

Interpretation of Findings and Comparison with Existing Literature

Discussion: Interpretation of Findings and Comparison with Existing Literature

This research investigates the application of supervised machine learning (ML) within two-stage matching models to address sample selection bias, a prevalent concern in strategy research (Heckman & Todd, 2009). The core contribution lies in guiding researchers on utilizing ML for selecting matching variables to enhance causal inference in propensity score matching (PSM) models. This approach is particularly relevant given the increasing emphasis on causal inference within the field (Bettis, Gambardella, Helfat, & Mitchell, 2014) and the growing availability of large-scale datasets with numerous interdependent variables (Reineke, Katila, & Eisenhardt, 2024; Katila, Piezunka, Reineke, & Eisenhardt, 2022).

The study demonstrates the ML-PSM technique using real-world technology invention data concerning public-private relationships. The findings suggest that ML offers a viable alternative to ad-hoc matching methods, potentially improving the precision and reproducibility of causal effect estimates (Angrist & Frandsen, 2022). This improvement stems from ML's ability to comprehensively evaluate a wide range of variables and their interactions, thereby acting as a data pre-processing method (Goller, Lechner, Moczall, & Wolff, 2020; Rathje & Katila, 2021). ML can identify and prioritize relevant confounders, prompting researchers to critically examine their underlying assumptions. Furthermore, ML automates the exploration of pairwise interactions and complex functional forms, which are often infeasible to handle manually (Mullainathan & Spiess, 2017).

The paper acknowledges that while ML methods are associative, they can incorporate a substantial number of observables, potentially mitigating unobserved heterogeneity. This aligns with the broader trend of leveraging AI and ML in strategy research, including comparing human and algorithmic strategies (Bahceci, Katila, & Miikkulainen, 2015), developing comprehensive search and innovation strategies (Bahceci, Katila, & Miikkulainen, 2023), and facilitating inductive theorizing through variable creation from unstructured data (Helfat et al., 2023; Miric, Jia, & Huang, 2022). The research also draws parallels with applications of ML in matching designs within other disciplines, such as economics (Angrist & Frandsen, 2022; Goller et al., 2020), medicine, and biosciences (D'Agostino, 1998; McCaffrey, Ridgeway, & Morral, 2004; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008; Cannas & Arpino, 2019; Lee, Lessler, & Stuart, 2010; Blakely, Lynch, Simons, Bentley, & Rose, 2020).

However, the study also emphasizes the limitations of ML-PSM. Supervised ML methods do not inherently provide standard errors, hindering direct interpretation of coefficient estimates (Mullainathan & Spiess, 2017). Moreover, the researcher's role in selecting the initial set of confounding variables remains crucial, highlighting the importance of human interpretation and domain expertise.

Interpretation of Findings and Comparison with Existing Literature

This research investigates the phenomenon of overreliance in human-AI decision-making, specifically addressing the limited success of explanations in mitigating this bias. Prior literature suggests that explanations either exacerbate or do not alter overreliance unless cognitive forcing functions are implemented to compel engagement with the task. This paper challenges this perspective, arguing that explanations *can* reduce overreliance under specific conditions, without the need for such interventions.

The central argument is that individuals strategically choose whether to engage with AI explanations based on a cost-benefit analysis. This framework posits that the cognitive costs associated with task completion and explanation verification are weighed against the perceived benefits of relying on the AI's prediction. The authors propose that previous studies showing null effects of explanations on overreliance may have inadvertently focused on scenarios where the cognitive costs of the task and the explanation are roughly equivalent and non-trivial. In such situations, individuals may opt to rely on the AI without thorough verification.

To test this framework, the authors conducted five studies (N=731) using a maze-solving task where participants collaborated with a simulated AI. Study 1 (N=340) manipulated task difficulty, replicating previous findings of equal overreliance levels between AI predictions alone and AI predictions with explanations for an easy task. However, as task difficulty increased, overreliance was significantly reduced in the explanation condition compared to the AI-only condition. This suggests that when the task is more effortful, the relative cognitive effort of engaging with the explanation becomes less, leading to a reduction in overreliance. Studies 2 (N=340) and 3 (N=286) further explored this by manipulating the cognitive cost of understanding the AI's explanations. The findings indicated that lower cognitive effort associated with explanations resulted in a greater reduction in overreliance.

Interpretation of Findings and Comparison with Existing Literature

This study investigates the impact of AI-powered Go programs (APGs) on the decision-making quality of professional Go players. The research compares human moves to AI's superior solutions before and after the public release of APGs, analyzing 750,990 moves in 25,033 games by 1,242 professional players from 2015 to 2019. The methodology involves using APG's best solutions as a benchmark to calculate the winning probability for each move, even for games played before APG's release.

The findings indicate a significant improvement in the quality of moves by professional Go players following the release of APGs. Before the release, the winning probability of each move by professional Go players averaged 2.47 percentage points lower than APG's moves. This gap decreased by approximately 0.756 percentage points (30.5 percent) after the release. This improvement in move quality correlates with an increased likelihood of winning the game. The effect is most pronounced in the early stages of the game, characterized by higher uncertainty. Furthermore, younger players, who are more adept at utilizing APGs, exhibit more significant improvements.

Mediation analysis reveals that improvements in move quality are primarily driven by a reduction in the number of errors (moves where the winning probability drops by 10 or more percentage points) and a decrease in the magnitude of the most critical mistake during the game. Specifically, the number of errors per game decreased by 0.15–0.50, and the magnitude of the most critical mistake decreased by 4–7 percentage points.

These findings contribute to the literature on AI and human capital by providing micro-level evidence of how AI can improve native human abilities in decision-making, even when the AI technology is not directly available during the task. This contrasts with studies focusing on AI as a direct assistant, highlighting AI's potential instructional role in training human professionals. The results also have implications for understanding digital literacy and inequality in AI utilization, as younger players benefit more from APGs.

Discussion: Interpretation of Findings and Comparison with Existing Literature

This paper investigates and compares decision-making processes for self-optimizing autonomic computing systems, focusing on heuristic, control-theoretical, and machine learning methods. The research addresses the challenge of designing the decision phase in autonomic systems, which is crucial for achieving self-configuring, self-healing, self-optimizing, and self-protecting capabilities. The study leverages the Application Heartbeats framework to provide sensor data and evaluates the performance of various decision mechanisms using standard benchmarks from the PARSEC suite.

The core contribution lies in the comparative analysis of these techniques, examining their properties and guarantees both theoretically and practically. The research synthesizes, develops, implements, and tests these techniques within a common framework, applying them to benchmark test cases. The results are presented in detail for a single benchmark and aggregated to summarize the effects observed across multiple case studies.

The paper acknowledges the potential of feedback control theory in autonomic systems, highlighting its underutilization despite its capacity to provide online time and convergence guarantees. A key challenge identified is the difficulty in developing accurate system models required for control-theoretical approaches. In contrast, machine learning techniques are presented as requiring less explicit modeling, as they can capture complex relationships online. The authors suggest that the most effective solutions may involve combining these techniques, enhancing feedback control with machine learning and vice versa.

The study also discusses the motivations for building autonomic computing systems, emphasizing self-management and its four key aspects: self-configuration, self-protection, self-healing, and self-optimization. While self-configuration is deemed beyond the scope of this work, the paper explores the role of decision-making in self-protection and self-healing. Machine learning techniques are highlighted for their utility in intrusion detection systems and failure classification, citing examples where decision trees, neural networks, and k-th nearest neighbor algorithms have been successfully applied to diagnose network problems. The authors posit that heuristic and control-theoretical approaches may be less effective for failure classification compared to machine learning.

Interpretation of Findings and Comparison with Existing Literature

This research addresses the gap in AI-assisted decision-making by proposing offline reinforcement learning (RL) as a method for optimizing human-AI interaction for diverse, human-centric objectives, moving beyond the predominant focus on decision accuracy. The study instantiated this approach with two objectives: human-AI accuracy and human skill improvement.

The methodology involved learning decision support policies from existing human-AI interaction data using offline RL. The problem of human-AI decision-making was framed as a Markov Decision Process, allowing for the derivation of optimal support policies. The reward function was designed to capture objectives like human learning, while the state space incorporated human-centric and contextual factors (e.g., human skill, AI uncertainty). The action space consisted of different types of human-AI interaction techniques.

Two experiments were conducted (N = 316 and N = 964), and the results indicated that policies optimized for accuracy led to significantly better accuracy and human-AI complementarity compared to other types of AI support. However, optimizing for human learning proved more challenging, with significant learning improvement observed only at times among participants interacting with learning-optimized policies.

The findings suggest that offline RL is a promising approach for modeling human-AI decision-making dynamics, leading to policies that can optimize human-centric objectives. The research emphasizes the importance of considering objectives beyond decision accuracy in AI-assisted decision-making, opening a novel research challenge of optimizing human-AI interaction for such objectives. The study also highlights the potential of dynamic AI support, which adapts based on context and objectives, to prevent overreliance on AI and encourage deeper cognitive engagement. This contrasts with existing decision support paradigms that often provide fixed types of support regardless of context.

Interpretation of Findings and Comparison with Existing Literature

This paper investigates the design and evaluation of human-centered Explainable AI (XAI) systems, focusing on how non-expert users perceive automatically generated rationales. The research builds upon the concept of interpretability, distinguishing it from explanation generation as a form of post-hoc interpretability. While interpretability focuses on understanding how a model works, explanation generation aims to provide useful information to users in an accessible manner. The study addresses the increasing complexity of AI systems, which are often perceived as "black boxes," particularly by non-expert users. The central argument is that explainability is crucial for building rapport, confidence, and understanding between AI agents and users, especially in the context of failures and unexpected behavior.

The methodology involves a case study using a modified version of the Frogger game, a sequential decision-making task. Participants engaged in a "think-aloud" protocol, verbalizing their reasoning for each action taken in the game. This process was divided into three phases: a guided tutorial, rationale collection, and transcribed explanation review. During rationale collection, an automatic speech-to-text library transcribed the utterances, reducing participant burden. The resulting corpus consisted of linked game states, actions, and explanations. Data was collected remotely from 60 participants, resulting in over 2000 action-explanation pairs.

A key component of the study is the use of an encoder-decoder recurrent neural network to generate natural language explanations for given actions. The study explores how the choice of words in the generated rationale affects human-factor dimensions such as confidence in the agent's decision, understandability, human-likeness, and explanatory power. The research highlights the importance of considering the data collection process, the algorithm's use of data, and the intended recipient of the explanation when evaluating the success of an XAI system. The authors emphasize the need for end-to-end studies incorporating fully-realized AI agents and explanation generation systems into human-subject studies to facilitate better design and evaluation of XAI systems.

Interpretation of Findings and Comparison with Existing Literature

This research article investigates the impact of Artificial Intelligence (AI) on decision-making processes within organizations, focusing on its consequences for individuals, companies, and society. The study highlights AI's role in analyzing data insights, identifying gaps, and transforming decision-making proficiency, particularly under time constraints. This aligns with existing literature (Duan, Edwards & Dwivedi, 2019; Roser, 2022) that acknowledges AI's growing presence and capabilities in handling large datasets.

The findings suggest that AI improves business prediction, transforms business strategies, and facilitates quick decision-making, leading to increased productivity and business success. This is achieved through automation efficiency and complex management, including handling multilayered problems, succession planning, and risk monitoring. These observations are consistent with previous studies (Favour Oluwadamilare Usman et al, 2024; Muhammad Eid BALBAA, 2024; Simon Kaggwa et al., 2023) that emphasize AI's ability to modernize processes, reduce timeframes, and increase responsiveness in decision-making. The research also points out that AI transforms internal operations across departments like transport and consultation management by predicting demands, adjusting supply levels, and optimizing results, thereby reducing operating costs.

Furthermore, the study acknowledges the challenges associated with AI implementation, including ethical concerns, biased algorithms, and social allegations. It emphasizes the importance of data privacy protection, transparency, and accountability in AI-driven decision-making. This aligns with existing literature (Muhammad Eid BALBAA, 2024) that highlights the need for precise data checking and authentication to avoid adverse effects from inadequate or influenced data. The research underscores the necessity of ethical AI methods, nonbiased algorithms, and governance regulations to ensure AI systems align with organizational aims and goals.

Discussion: Interpretation of Findings and Comparison with Existing Literature

This research investigated the phenomenon of overreliance on AI in AI-assisted decision-making, addressing the limitations of explainable AI (XAI) in mitigating this issue. The central premise, informed by the dual-process theory of cognition, posits that individuals often rely on heuristic-based System 1 thinking, rather than engaging in analytical System 2 processing when interacting with AI explanations. This leads to the development of generalized trust heuristics regarding the AI's competence, rather than a critical evaluation of individual explanations. Consequently, the study explored the efficacy of cognitive forcing functions, interventions designed to disrupt heuristic reasoning and promote analytical engagement with AI-generated explanations.

The experimental results (N=199) demonstrated that cognitive forcing functions significantly reduced overreliance on AI compared to simple XAI approaches and a no-AI baseline. This finding supports the hypothesis that interventions promoting analytical thinking can improve human-AI collaboration by reducing the tendency to blindly accept AI suggestions. However, the study also revealed a trade-off: interventions that most effectively reduced overreliance were perceived as less usable and received less favorable subjective ratings. This suggests that while cognitive forcing functions can be effective, their design must carefully consider user experience to ensure acceptability and sustained engagement.

Furthermore, the research examined potential intervention-generated inequalities by analyzing the impact of cognitive forcing functions on individuals with varying levels of Need for Cognition (NFC). The findings indicated that these interventions disproportionately benefited participants with higher NFC, suggesting that individual cognitive motivation moderates the effectiveness of XAI solutions. This highlights the importance of considering individual differences in cognitive style and motivation when designing and deploying AI-assisted decision-making tools. The study's findings align with existing literature suggesting that explanations can be interpreted as general signals of competence, increasing trust and overreliance on AI, rather than being individually evaluated. The research extends this understanding by demonstrating that cognitive forcing functions can partially counteract this effect, albeit with potential usability trade-offs and differential impacts based on individual cognitive traits.

Discussion: Interpretation of Findings and Comparison with Existing Literature

This research challenges the prevailing view in the literature that explanations provided by AI systems fail to reduce overreliance in human-AI decision-making. Prior studies have suggested that explanations either exacerbate or do not alter overreliance, often attributing this to cognitive biases or uncalibrated trust, necessitating interventions like cognitive forcing functions to compel engagement. However, this work demonstrates that explanations *can* reduce overreliance without such interventions, contingent on the cognitive costs and benefits associated with engaging with the task and the explanation itself.

The core argument is that previous research has often focused on scenarios where the cognitive cost of verifying an AI's explanation is comparable to the cost of completing the task independently. In such situations, individuals may opt to rely on the AI's prediction without thorough verification. To address this, the authors propose a cost-benefit framework wherein individuals strategically weigh the cognitive effort required to complete the task alone or verify the AI's explanation against the seemingly effortless strategy of overreliance. This framework posits that explanations will reduce overreliance when tasks are more effortful, and the explanations sufficiently reduce the cost of verification, or when the explanations are inherently easier to understand.

This theoretical framework was empirically tested through five studies (N=731) using a maze-solving task. Study 1 manipulated task difficulty, replicating prior findings of equal overreliance levels with and without explanations for an easy task. Critically, as task difficulty increased, overreliance decreased with explanations compared to AI predictions alone. Studies 2 and 3 further explored the cognitive cost of understanding explanations, finding that lower cognitive effort led to a greater reduction in overreliance. Study 4 examined the balance between cognitive effort (costs) and incentives (benefits) in decision-making. Study 5 adapted the Cognitive Effort Discounting paradigm to quantify the utility of different explanations, providing further support for the proposed cost-benefit framework.

These findings suggest that the null effects observed in previous literature may be attributable to explanations not sufficiently reducing the cognitive costs of verifying the AI's prediction. By manipulating task difficulty and explanation complexity, this research demonstrates that explanations can indeed reduce overreliance when they offer a more efficient means of arriving at a correct decision compared to both independent problem-solving and blind reliance on the AI. This work contributes a nuanced understanding of the conditions under which XAI can effectively improve human-AI collaboration and mitigate the risks associated with overreliance.

Interpretation of Findings and Comparison with Existing Literature

This commentary introduces initial concepts for a draft Code of Conduct framework designed for the development and application of artificial intelligence (AI) in health, health care, and biomedical science. The increasing deployment of digital tools since the 1970s has led to an explosion of health data, with the aim of leveraging this data to transform health outcomes through a continuously learning health system (LHS) (National Academy of Medicine, 2022; IOM, 2007). While progress has been incremental through expert systems, clinical decision support, and machine learning, the translation of AI prototypes into practice has been limited.

The National Academy of Medicine (NAM) previously highlighted AI's potential to disrupt and transform health care, augmenting human capacity and improving health, while also acknowledging risks to equity, safety, and privacy (Matheny et al., 2022). The commentary notes the rapid evolution of advanced predictive and generative AI, including large language models (LLMs) like ChatGPT, necessitating rapid adaptation and alignment on guardrails for responsible AI use (Hutson, 2022). This aligns with the LHS principles of safe, effective, patient-centered, timely, efficient, and equitable care, further emphasizing transparency, accountability, and security (IOM, 2000; IOM, 2001).

The commentary acknowledges that AI methods are being employed in various applications, including disease screening, outcome prediction, personalized treatment, and drug discovery (Ardila et al., 2019; Cai et al., 2015; Glover et al., 2022; Gulshan et al., 2016; Rajpurkar et al., 2022). However, AI outputs, built on human-influenced datasets and models, may result in biases, particularly for under-represented groups (Obermeyer et al., 2019; Christensen et al., 2021). Risks include misdiagnosis, resource overuse, privacy breaches, and workforce displacement (Matheny et al., 2022). The evolving nature of AI models can also contribute to user confusion.

Numerous frameworks from intergovernmental agencies, private enterprises, and academic communities guide responsible AI use. The Asilomar AI Principles (Future of Life Institute, 2017) advanced guidelines covering research priorities, safety considerations, and value alignment. However, a need remains for convergence and interoperability (Dorr et al., 2023). The commentary suggests that the granularity of existing frameworks may hinder adoption, particularly given the challenges of applying principles like explainability and reproducibility to LLMs.

The proposed AI Code of Conduct framework comprises harmonized principles and simple rules (Code Commitments) for stakeholders, including developers, researchers, health systems, payers, patients, and federal agencies. The Code Principles aim to shape health AI governance, while the Code Commitments, building from LHS principles, are intended to direct the application and evaluation of the Code Principles. This approach is consistent with complex adaptive systems (CAS) theory, where a small set of rules can create desired outcomes at the system level (IOM, 2001).

Limitations of the Study

Discussion: Limitations of the Study

This study, while illuminating the Social Security Administration's (SSA) pioneering use of Artificial Intelligence (AI) in adjudication, possesses inherent limitations that warrant consideration.

The research primarily focuses on a single case study: the SSA Disability Program. While this allows for an in-depth analysis of the agency's AI implementation, the generalizability of the findings to other adjudicative systems or government agencies may be limited. The specific organizational structure, data infrastructure, and policy environment of the SSA may not be directly transferable to other contexts.

Furthermore, the study acknowledges reliance on prior work by some of the co-authors, potentially introducing a degree of confirmation bias or limiting the scope of inquiry to previously established findings. The research also does not explicitly address potential biases embedded within the AI algorithms themselves or the data used to train them, which could perpetuate existing inequalities in disability determinations. While the study highlights the importance of "blended expertise," it does not delve into the challenges of recruiting, training, and retaining personnel with the necessary interdisciplinary skills. Finally, the study's conclusions are based on the SSA's experience up to the point of publication; the long-term impacts of AI implementation on adjudicatory accuracy, consistency, and speed, as well as potential unintended consequences, remain to be fully assessed.

Limitations of the Study

This policy brief acknowledges several limitations inherent in the application of Artificial Intelligence (AI) to military decision-making. A primary limitation stems from the brittleness and unreliability of even advanced Machine Learning (ML) systems. These systems, designed to identify patterns in data for predictive purposes, often exhibit performance degradation or failure when confronted with unforeseen environmental changes or adversarial interference. Current techniques aimed at enhancing robustness are not sufficiently mature to guarantee reliable operation, particularly under the dynamic and hostile conditions of warfare. The continuous learning and updating processes required for deployed systems further exacerbate this challenge, potentially introducing vulnerabilities and flaws with incompletely understood consequences.

Furthermore, the study recognizes the competitive pressures militaries face to adopt ML systems, despite their inherent uncertainties. The perceived need to maintain parity or superiority with potential adversaries, such as China, may incentivize the rapid deployment of AI capabilities with unproven reliability. This risk is particularly acute in seemingly benign applications supporting human decision-making or in non-lethal roles like cyber operations. The potential for flawed ML capabilities to be deployed on both sides of a U.S.-China confrontation introduces additional complexities and uncertainties.

Limitations of the Study

This paper provides an overview of efforts to resolve the gap between AI principles and practice, presenting a typology and catalog of 35 recent efforts to implement AI principles and exploring three case studies in depth. While the selected case studies (Microsoft's AETHER Committee, OpenAI's Staged Release of GPT-2, and the OECD AI Policy Observatory) were chosen for their scope, scale, and novelty, the study's focus on these specific examples inherently limits the generalizability of its findings. The lessons derived from these cases may not be directly transferable to all organizations or contexts involved in AI development and deployment.

The research methodology relies on case study analysis, which, while providing in-depth insights, may not capture the full complexity of the AI governance landscape. The paper acknowledges the existence of at least 84 AI principles and strategy documents as of September 2019, but the analysis is confined to a subset of 35 efforts and three detailed case studies. This selective approach may introduce bias, as the chosen examples might not be representative of the broader range of initiatives and challenges in AI governance.

Furthermore, the study's conclusions are based on data and observations up to the point of publication. Given the rapid pace of technological advancement and the evolving nature of AI governance, the findings may not fully reflect the current state of the field. The paper acknowledges the potential for "technological, organizational, and regulatory lock-in effects," suggesting that the initial efforts analyzed may have a disproportionate influence. However, the long-term impacts and sustainability of these efforts remain uncertain and warrant further investigation.

Limitations of the Study

This working paper acknowledges several limitations inherent in the current discourse and practical application of Artificial Intelligence (AI), shaping the scope and conclusions of the analysis.

First, the paper emphasizes that Machine Learning (ML), the dominant technique underpinning contemporary AI, is fundamentally a context-dependent statistical inference method. This inherent characteristic restricts the applicability of current AI systems to well-specified problem domains, limiting their capacity for generalization and contextual understanding compared to human cognition. The study cautions against extrapolating from successes in narrow AI applications to broader claims of artificial general intelligence (AGI).

Second, the analysis highlights the interconnectedness of AI with other digital technologies, particularly digital platforms and big data infrastructure. The paper argues that focusing solely on AI governance, without considering the broader ecosystem of intelligent tools and systems, provides an incomplete and potentially distorted view of the challenges. The study suggests that the regulation of digital platforms and big data is intrinsically linked to governing AI, as these platforms generate the data upon which AI tools operate.

Third, the paper identifies the difficulty in translating abstract objectives, such as digital privacy and transparency, into concrete operational terms. The authors acknowledge that achieving consensus on specific goals and implementation strategies is a significant challenge, particularly across diverse communities and international contexts.

Finally, the study recognizes the potential for overstating the economic implications of AI applications. The authors question whether public investment should prioritize basic research or focus on the diffusion of AI tools, user interfaces, and the training needed for their effective implementation. This highlights a limitation in current understanding regarding the optimal allocation of resources for AI development and deployment.

Limitations of the Study

This study, while demonstrating the potential of cognitive forcing functions to mitigate overreliance on AI in decision-making, acknowledges several limitations. Firstly, despite the significant reduction in overreliance compared to simple explainable AI approaches, cognitive forcing functions did not entirely eliminate the phenomenon. Participants, even under cognitive forcing conditions, still exhibited a tendency to rely on incorrect AI predictions in scenarios where they would have made better decisions independently. Secondly, the research identified a trade-off between the effectiveness of the interventions and their acceptability. Designs that most effectively reduced overreliance were perceived as more difficult, less preferred, and less trusted by the participants. Finally, the study revealed that cognitive forcing interventions disproportionately benefited individuals with a higher Need for Cognition (NFC), suggesting that the effectiveness of these interventions is moderated by individual cognitive motivation. This indicates a potential for intervention-generated inequalities, where individuals with lower NFC may not experience the same benefits from cognitive forcing functions as those with higher NFC. The findings highlight the complex interplay between intervention design, user perception, and individual cognitive traits in AI-assisted decision-making.

Limitations of the Study

This research is subject to certain limitations. The empirical experiment, while providing insights into user responses to explanations, particularly among novice users with low self-competence, revealed that explanations do not invariably improve trust calibration and can, in some instances, be detrimental. This suggests that the efficacy of explanations is contingent upon user characteristics and potentially other contextual factors not fully explored in this study. Furthermore, the study acknowledges the challenge of translating research findings into tangible guidance for industry designers, highlighting the need for a more comprehensive and accessible educational resource. The exploratory interviews and usability testing conducted with industry designers, while valuable, may not fully represent the diverse perspectives and needs of all practitioners in the field. The research also recognizes the limitations associated with "black box" models and the protection of intellectual property, which can hinder the development and implementation of effective explainable interfaces.

Discussion: Limitations of the Study

While this research provides evidence that explanations can reduce overreliance on AI systems under specific conditions, several limitations warrant consideration. The study primarily employed a maze-solving task, which, despite its utility in manipulating cognitive costs, may not fully generalize to more complex, real-world decision-making scenarios. The artificial nature of the task and the simulated AI may not capture the nuances of human-AI interaction in applied settings, where trust, risk perception, and domain expertise play a more significant role.

Furthermore, the cost-benefit framework, while providing a useful theoretical lens, simplifies the multifaceted cognitive processes underlying human reliance on AI. The model assumes a rational actor weighing costs and benefits, potentially overlooking the influence of cognitive biases, emotional factors, and individual differences in decision-making styles. The quantification of cognitive effort, particularly in Study 5 using the Cognitive Effort Discounting paradigm, presents inherent challenges in accurately capturing subjective experiences.

The manipulation of explanation difficulty in Studies 2 and 3, while demonstrating the impact of cognitive cost, may not fully represent the spectrum of explanation types and their varying levels of comprehensibility. The study's focus on specific types of explanations may limit the generalizability of the findings to other explanation modalities, such as counterfactual explanations or feature importance visualizations.

Finally, the sample size of Study 4 (N=114), which examined the interplay between cognitive effort and monetary incentives, is relatively small compared to the other studies. This smaller sample size may limit the statistical power to detect subtle effects and could potentially impact the robustness of the conclusions drawn from this particular experiment. Future research should address these limitations by exploring the generalizability of the findings across diverse tasks, explanation types, and real-world contexts, while also incorporating more nuanced models of human cognition and decision-making.

Limitations of the Study

This dissertation investigates the vulnerabilities of data-based decision-making (DBDM) systems, focusing on challenges arising from data reliance. The research explores these vulnerabilities through three distinct perspectives: adversarial learning, game theory, and online reinforcement learning (RL).

One limitation lies in the scope of the adversarial learning investigation, which primarily focuses on end-to-end poisoning attacks on learning-planning models. While the study demonstrates the feasibility of such attacks, it does not exhaustively cover all potential attack vectors or defense mechanisms. The research findings are specific to the attack methodologies employed and the model architectures tested, potentially limiting the generalizability of the conclusions to other DBDM systems.

The game theory analysis, centered on security games, investigates how a DBDM system can form a behavioral model despite malicious manipulation of observable data. A limitation here is the reliance on specific game-theoretic models, such as Stackelberg Security Games (SSGs), and human behavior models like Quantal Response (QR) and Subjective Utility Quantal Response (SUQR). The accuracy and applicability of these models in representing real-world attacker behavior may vary, introducing potential biases in the analysis. Furthermore, the study's focus on partial behavior deception models and the Cognitive Hierarchy approach may not fully capture the complexity of strategic interactions in dynamic environments.

The examination of online RL settings highlights the challenges of limited and sparse data in building effective DBDM models. A limitation in this area is the potential sensitivity of RL algorithms to hyperparameter tuning and exploration strategies. The effectiveness of the proposed solutions may depend on the specific characteristics of the environment and the reward structure, which could limit their applicability across diverse real-world scenarios.

In summary, while this dissertation provides valuable insights into the vulnerabilities of DBDM systems, the findings are subject to limitations related to the scope of adversarial attacks considered, the reliance on specific game-theoretic and behavioral models, and the challenges of generalizing RL-based solutions across different environments. Further research is needed to address these limitations and develop more robust and resilient DBDM systems.

Limitations of the Study

This study primarily focuses on the theoretical application of regulatory design principles to address the challenges of AI decision-making in government, particularly concerning vulnerable populations. The analysis is largely based on existing literature and case studies, notably the "Robodebt" scheme in Australia, to illustrate the potential pitfalls of poorly designed AI systems. While the proposed regulatory innovations, including the "Golden Rule" and a monitoring body to ensure well-calibrated decisions, offer a novel framework, their practical implementation and effectiveness remain to be empirically tested. The article acknowledges the necessity of managing conflicting norms, such as efficiency and fairness, but the specific mechanisms for balancing these competing interests in real-world scenarios are not fully explored. Furthermore, the study's reliance on a systems perspective, while valuable for holistic analysis, may not fully capture the nuances and complexities of individual cases and the diverse needs of vulnerable populations affected by AI-driven decisions. The generalizability of the findings may also be limited by the specific context of the Australian legal and administrative system.

Limitations of the Study

This study, while demonstrating the utility of machine learning (ML) in propensity score matching (PSM) for causal inference, acknowledges several limitations. First, supervised machine learning methods, as implemented in this context, do not inherently provide standard errors, thus precluding direct interpretation of coefficient estimates (Mullainathan & Spiess, 2017). This absence of readily interpretable coefficients necessitates caution when drawing definitive conclusions based solely on the ML-PSM results.

Second, the study emphasizes that the application of ML techniques does not absolve researchers of their responsibility in the research process. The selection of the initial set of confounding variables, which forms the basis for the ML algorithm's pruning process, remains a critical task requiring human judgment and domain expertise. The quality and relevance of these initial variables directly impact the effectiveness of the ML-PSM approach. Furthermore, the automated nature of confounder selection does not eliminate the need for researchers to critically evaluate and interpret the variables chosen by the algorithm.

Finally, while the study demonstrates the ML-PSM technique using a technology invention dataset, the focus remains on methodological demonstration rather than substantive findings. Although the second-stage results are reported and found to be consistent across different matching methods, the primary objective is to illustrate the application and potential of ML in addressing sample selection bias. Therefore, the generalizability of the specific findings related to public funding impact in technology strategy should be interpreted with caution.

Implications for Future Research and Practice

Implications for Future Research and Practice

This research demonstrates the potential of offline reinforcement learning (RL) as a method for modeling human-AI decision-making, leading to policies that optimize human-centric objectives. The findings emphasize the importance of considering objectives beyond decision accuracy in AI-assisted decision-making, thereby opening a novel research challenge: optimizing human-AI interaction for such objectives. The study's methodology involved casting human-AI decision-making as a Markov Decision Process and learning policies via offline RL to optimize different objectives. The approach is customizable, allowing for the optimization of various objectives by crafting the reward function and constructing appropriate state and action spaces. Two experiments (N = 316 and N = 964) consistently showed that policies optimized for accuracy led to significantly better accuracy and human-AI complementarity compared to other AI support types. However, optimizing human learning proved more challenging, with significant learning improvement observed only at times among participants interacting with learning-optimized policies. Future research should explore the optimization of a broader range of human-centric objectives, such as skill improvement, autonomy, and social belonging, within AI-assisted decision-making contexts. Further investigation is also needed to refine the methodology for optimizing human learning in these settings.

Implications for Future Research and Practice

This chapter underscores the necessity for continuous dialogue and research to ascertain the role of ChatGPT and other AI-driven tools in shaping the trajectory of global higher education. The authors advocate for discussions and research endeavors that navigate ethical considerations, technological advancements, and the broader implications of AI in education. These efforts should involve diverse stakeholders to advance the responsible and beneficial integration of AI into global higher education, while simultaneously mitigating potential harm and misuse. The chapter highlights the importance of disentangling the complex interplay of opportunities and challenges presented by AI, emphasizing the role of scholars in this process.

The study identifies several key areas for future investigation, including data privacy, biases, and ethical considerations related to the integration of ChatGPT into academics. Furthermore, the readiness of faculty and students to adopt and effectively utilize AI tools is highlighted as a critical area for further research. The chapter proposes recommendations to address these challenges and foster effective AI utilization, suggesting a need for empirical studies to evaluate the efficacy of these recommendations.

The authors' engagement with ChatGPT, leveraging its generative capabilities, has enabled personalized learning experiences, stimulated critical thinking, and encouraged creative problem-solving among students. This experience underscores the potential of AI to cultivate a more interactive, inclusive, and intellectually stimulating academic environment. The chapter advocates for further experimentation and open-mindedness in embracing emerging technologies that can revolutionize knowledge dissemination and acquisition. The aim is to ensure that AI serves as a catalyst for innovation, inclusivity, and intellectual growth within higher education.

Implications for Future Research and Practice

This research provides a cost-benefit framework for understanding and predicting the impact of AI explanations on human overreliance. The central finding is that explanations can reduce overreliance when the cognitive cost of engaging with the explanation is sufficiently lower than the cost of performing the task independently. This contrasts with prior work suggesting that explanations often fail to reduce or even exacerbate overreliance.

The study employed a series of five experiments (N=731) using a maze-solving task where participants collaborated with a simulated AI. The methodologies involved manipulating task difficulty (Study 1), explanation difficulty (Studies 2 and 3), and monetary compensation (Study 4) to assess their impact on overreliance. Study 5 adapted the Cognitive Effort Discounting paradigm to quantify the utility of different explanations.

Key conclusions include: (1) Task difficulty influences overreliance, with explanations becoming more effective at reducing overreliance as the task becomes more cognitively demanding. (2) Explanation difficulty also plays a crucial role; simpler, easier-to-understand explanations lead to a greater reduction in overreliance. (3) Individuals strategically weigh the costs and benefits of engaging with explanations versus relying solely on the AI or performing the task independently. The findings suggest that the relative cognitive effort required to verify an AI's explanation, compared to the effort of completing the task manually, is a critical determinant of whether explanations will reduce overreliance.

Implications for Future Research and Practice

This research highlights the potential of cognitive forcing functions as a means to mitigate overreliance on AI in AI-assisted decision-making contexts. The study (N=199) compared three cognitive forcing designs with two explainable AI approaches and a no-AI baseline, revealing that cognitive forcing significantly reduced overreliance compared to simple explainable AI methods. These findings suggest that interventions prompting analytical engagement with AI-generated explanations can be more effective than explanations alone in improving human-AI team performance.

However, the study also identified a critical trade-off: interventions that most effectively reduced overreliance were perceived as less usable and received less favorable subjective ratings. This implies that future research should focus on designing cognitive forcing functions that are both effective in promoting analytical thinking and acceptable to users. Further investigation is needed to understand the specific design elements that contribute to usability and acceptability, while maintaining the cognitive forcing effect.

Furthermore, the research revealed that the benefits of cognitive forcing interventions were not uniformly distributed. Participants with a higher Need for Cognition (NFC) benefited more from these interventions, suggesting that individual cognitive motivation moderates the effectiveness of explainable AI solutions. This finding underscores the importance of considering individual differences in cognitive styles and motivations when designing AI-assisted decision-making systems. Future research should explore adaptive interventions that tailor the level of cognitive forcing to individual user characteristics, potentially maximizing the benefits for a broader range of users.

In practice, these findings suggest that implementing cognitive forcing functions in real-world AI-assisted decision-making tools could improve decision quality by reducing overreliance on AI. However, careful consideration must be given to the design of these interventions to ensure they are user-friendly and do not disproportionately benefit individuals with high NFC. Future practical applications should incorporate user feedback and iterative design processes to optimize the balance between effectiveness and acceptability.

Implications for Future Research and Practice

The research findings suggest that large language models (LLMs), such as GPT-3, possess the capacity to generate persuasive public health messages, specifically pro-vaccination messages, that can be perceived as more effective and evoke more positive attitudes compared to human-authored messages from institutions like the CDC. Study 2 demonstrated that curated GPT-3-generated messages were perceived as stronger arguments and more effective than CDC messages. However, Study 3 revealed a preference for messages originating from human institutions rather than AI sources when the source was explicitly labeled.

These findings have implications for future research and practice in public health communication. Firstly, further investigation is warranted to explore the optimal integration of LLMs into public health messaging workflows. The study highlights the potential for LLMs to augment activities such as content development by rapidly generating tailored messages, drawing upon extensive verbal user data. However, the non-deterministic nature of LLMs and the potential for generating inaccurate or inappropriate content necessitate human supervision and quality control.

Future research should focus on developing best practices for assessing the generative outputs of LLMs in social science research, as well as methods for health professionals to effectively utilize AI systems to augment public health messaging. This includes exploring strategies for mitigating the negative perceptions associated with AI-generated content through careful source attribution and transparency. Methodologically, the research underscores the importance of evaluating both the persuasive quality and the perceived source credibility of AI-generated messages. The study employed a multi-method approach, including systematic evaluation of GPT-3's message generation capabilities (Study 1), comparative analysis of AI-generated and human-authored messages (Study 2), and assessment of the impact of source labels on perceived quality (Study 3). This approach provides a framework for future research to rigorously assess the potential and limitations of LLMs in public health communication.

Implications for Future Research and Practice

This research highlights several avenues for future investigation and practical application concerning AI-driven decision-making in management. The study underscores the potential for AI to transform internal operations across various departments, including transport and consultation management, by predicting demands, adjusting supply levels, and optimizing results, leading to reduced operating costs. However, it also identifies critical challenges that warrant further attention.

Future research should focus on addressing ethical concerns, biases in AI algorithms, and social implications arising from the increased use of AI in decision-making. Specifically, investigations into data privacy protection, transparency, and accountability in AI-driven processes are crucial. The explainability of AI decisions remains a significant issue, impacting a company's reputation and requiring careful consideration.

Furthermore, the study suggests a need for research into the optimal balance between human and AI collaboration in decision-making. Understanding how to effectively integrate human oversight and judgment with AI capabilities is essential for mitigating risks and ensuring ethical outcomes.

From a practical standpoint, organizations should prioritize the development and implementation of ethical AI frameworks. These frameworks should emphasize transparency, fairness, and non-biased algorithms. Governance regulations and policies are necessary to address biases and ensure that AI systems align with organizational aims and goals. Further practical implications include the need for continuous monitoring and assessment of AI systems to identify and rectify potential issues related to data quality, algorithm performance, and unintended consequences. The study opens the scope for improvement in AI functionality, suggesting that ongoing research and development efforts should focus on enhancing the accuracy, reliability, and ethical soundness of AI-driven decision-making tools.

Implications for Future Research and Practice

The One Hundred Year Study on Artificial Intelligence, initiated in 2014, aims to provide expert-informed guidance for AI research, development, and systems design, as well as programs and policies to ensure broad societal benefits. The study is structured as a series of expert panel assessments conducted every five years. The inaugural report, focusing on "AI and Life in 2030" in a typical North American city, acknowledges that AI's impacts are interconnected with other societal and technological developments. The study's findings are intended to inform resource allocation in industry, governmental planning, and prioritization within the AI research community. Future studies are planned to expand the project's scope internationally. The report's limitations, including its focus on North American cities and exclusion of military applications, are acknowledged, emphasizing the need for continued monitoring and deliberation on AI's implications for defense and warfare. The study anticipates that future assessments will reflect on the successes and failures of this initial report in predicting AI's trajectory and influences.

Implications for Future Research and Practice

This research provides several avenues for future investigation and practical application in the realm of organizational decision-making and participatory AI design. The study demonstrates the potential of using ML models, not as automated decision-makers, but as boundary objects to facilitate stakeholder deliberation and reflection on existing decision-making processes. This approach highlights the value of stakeholder-centered participatory AI design, where individuals directly involved in or affected by decisions actively participate in the creation and evaluation of ML models.

The findings suggest that future research should explore the design of AI tools that specifically support deliberation and reflection, focusing on how these tools can help stakeholders identify and address biases, non-inclusive practices, and lack of diverse representation within their organizations. The "Deliberating with AI" web tool, described in this paper, serves as a prototype for such tools, demonstrating how stakeholders can create and evaluate ML models to examine the strengths and shortcomings of past decision-making.

Methodologically, the study employs a case study approach, applying the "Deliberating with AI" tool to the context of graduate school admissions. This provides valuable insights into how faculty and students can use ML models to deliberate over organizational decision-making practices. Future research could expand on this by applying the tool to other organizational contexts and decision-making scenarios, such as hiring, resource allocation, or project selection. Furthermore, longitudinal studies could assess the long-term impact of using such tools on organizational culture and decision-making outcomes.

The study's conclusion emphasizes that the goal is not to create socio-technical systems objectively devoid of biases, but rather to provide a method for users to deliberate on how to improve future personal and organizational decision-making. This perspective suggests that future research should focus on developing methods for evaluating the effectiveness of deliberation-focused AI tools, considering not only quantitative metrics such as accuracy or fairness, but also qualitative measures such as stakeholder engagement, shared understanding, and the generation of actionable insights. The research also calls for further exploration of how data-driven reflection can be used to help individuals and groups assess data regarding others, extending beyond its current application in personal health behavior insights.

Implications for Future Research and Practice

This study provides empirical evidence that AI can improve human decision-making capabilities, even when AI technology is not directly utilized during the task. Analyzing 750,990 moves in 25,033 professional Go games played by 1,242 players between 2015 and 2019, the research demonstrates a significant improvement in the quality of moves following the public release of AI-powered Go programs (APGs). The methodology involved comparing human moves to the APG's optimal solutions, using the change in winning probability as a metric for move quality.

Key findings indicate that the gap between human and AI move quality decreased by 0.756 percentage points (30.5%) after the release of APGs. This improvement was particularly pronounced in the early stages of games, characterized by higher uncertainty. Furthermore, younger players exhibited greater improvements, suggesting a generational disparity in AI adoption and utilization.

Mediation analysis revealed that the enhanced move quality was primarily driven by a reduction in the number of errors (moves with a significant drop in winning probability) and a decrease in the magnitude of the most critical mistake made during the game. Specifically, the number of errors per game decreased by 0.15–0.50, and the magnitude of the most critical mistake decreased by 4–7 percentage points. These findings suggest that AI can serve an instructional role, training human professionals to make better decisions by refining their heuristics and reducing critical errors.

Implications for Future Research and Practice

The study of the Social Security Administration's (SSA) pioneering use of Artificial Intelligence (AI) in adjudication offers several implications for future research and practice in administrative adjudication, AI governance, and public administration.

Key Findings and Methodologies:

- **AI Deployment in Adjudication:** The SSA has successfully developed and deployed AI systems to assist adjudicators in making core decisions, specifically in the Disability Program. This includes an AI system that checks draft decisions for approximately 30 quality issues, addressing accuracy, consistency, and speed of case processing.
- **Overcoming Organizational Barriers:** The SSA overcame significant organizational roadblocks to implement AI use cases.
- **Strategic Investments:** Early strategic investments in data infrastructure, policy, and personnel laid the groundwork for AI implementation.
- **Blended Expertise:** The importance of "blended expertise," which combines technical and domain knowledge, was crucial for identifying, developing, and testing AI innovations.
- **Continuous Improvement:** AI should be viewed as part of a continuous improvement process rather than a one-off IT product. AI governance is essentially quality assurance.

Implications for Future Research:

- **AI in Mass Adjudication:** Further research is needed to understand how AI can be developed in large organizations involved in mass adjudication, such as immigration, Medicare appeals, veterans benefits, and patent examination.
- **AI and Due Process:** The SSA case study suggests that AI can be deployed to advance due process goals in adjudication, warranting further investigation into how AI can enhance fairness and transparency in adjudicative processes.
- **Public Administration and Innovation:** The study confirms the importance of the location of the core organization responsible for innovation and the involvement of end-users in system development. Further research could explore the impact of AI systems on public administration, particularly in complex, discretion-laden areas.

Implications for Practice:

- **AI Governance Principles:** The SSA case offers practical lessons for implementing AI governance principles in the public sector, particularly in contested areas.
- **Leadership Support:** Leadership support is crucial for overcoming organizational barriers to AI adoption.
- **Operational Data:** The value of operational data in developing and deploying AI systems is highlighted.
- **Early Piloting and Continuous Evaluation:** Early and frequent testing, along with continuous iteration and evaluation, are essential for successful AI implementation.
- **Development and Deployment Ecosystem:** Establishing a robust development and deployment ecosystem is critical for AI innovation.

Conclusion

Summary of Key Findings

Conclusion: Summary of Key Findings

This paper addresses the critical gap between the articulation of AI principles and their practical implementation. It provides a typology and catalog of 35 recent efforts aimed at operationalizing AI principles, further exploring three case studies to offer guidance for stakeholders navigating the responsible development and deployment of AI.

Key Findings from Case Studies:

- **Microsoft's AETHER Committee:** This case study highlights the integration of AI principles into a major technology company's practices and policies. The AETHER Committee, composed of interdisciplinary working groups, addresses emerging questions related to bias, fairness, reliability, safety, and potential threats to human rights. Key drivers of success include executive and employee buy-in, integration into a broader company culture of responsible AI, and the creation of interdisciplinary working groups.

- **OpenAI's Staged Release of GPT-2:** This case examines responsible publication norms for advanced AI technologies. OpenAI's staged release of the GPT-2 language model, accompanied by threat modeling and documentation, aimed to promote transparency about potential risks and facilitate open communication with users to identify and mitigate harms.

- **The OECD AI Policy Observatory:** This case demonstrates how an intergovernmental initiative can facilitate international coordination in implementing AI principles. The OECD AI Policy Observatory provides practical guidance, maintains a database of AI policies and initiatives, compiles metrics on global AI development, and convenes stakeholders from the private sector, governments, academia, and civil society. Key factors for success include sustained government support, multistakeholder expert groups, and extensive outreach efforts.

Overarching Conclusion:

The operationalization of AI principles represents a critical juncture for AI stakeholders. Early efforts are particularly influential due to technological, organizational, and regulatory lock-in effects. The case studies analyzed in this report provide meaningful examples and lessons for stakeholders seeking to develop and deploy trustworthy AI technologies.

Conclusion: Summary of Key Findings

This report, commissioned by the Administrative Conference of the United States (ACUS), investigates the utilization of Artificial Intelligence (AI) within federal administrative agencies. The research, encompassing a rigorous survey of 142 federal entities, in-depth case studies of seven agencies, and cross-cutting analyses, yields five principal findings.

First, AI adoption is widespread across the federal administrative landscape. Approximately 45% of the agencies surveyed have experimented with AI and machine learning (ML) tools. These tools are applied across a spectrum of governance functions, including regulatory enforcement, adjudication of benefits, risk monitoring, data extraction, and public communication. The AI techniques employed range from conventional machine learning to advanced deep learning methodologies.

Second, despite the broad adoption, the sophistication of AI tools deployed by agencies is generally low. Evaluations indicated that only 16% of the AI applications were rated as highly sophisticated, raising concerns about the potential for a widening technology gap between the public and private sectors.

Third, the implementation of AI introduces significant accountability challenges. The inherent opacity of advanced AI tools complicates the requirement for public officials to provide reasoned explanations for decisions affecting public rights. Addressing this challenge necessitates context-specific approaches to transparency, adaptation of existing administrative law principles, and prospective benchmarking of AI tools against human decision-making.

Fourth, the development and responsible use of AI within agencies depend on internal technical capacity. A majority (53%) of the AI applications studied were developed in-house, highlighting the importance of cultivating internal expertise. In-house expertise facilitates the creation of AI tools tailored to complex governance tasks and ensures compliance with legal and policy requirements. Sustained investment in human capital and infrastructure is crucial for fully leveraging AI's potential.

Fifth, the increasing use of AI by government agencies carries the risk of exacerbating distributive inequalities and fueling political anxieties. Algorithmic predictions may disproportionately impact smaller entities lacking the resources to effectively navigate AI systems. Public trust in government may erode if AI systems are perceived as biased or manipulated.

Conclusion: Summary of Key Findings

The 2024 survey of data and analytics professionals highlights key objectives for contemporary data programs, notably advancing data-driven decision-making, controlling costs, and modernizing data ecosystems. Survey results indicate that organizations are prioritizing investments in data governance frameworks and location intelligence to bolster data capabilities and enhance operational efficiency.

Despite these advancements, significant challenges persist, particularly concerning data quality and overall trust in data. A substantial proportion of organizations report ongoing difficulties in ensuring data accuracy, reliability, and security, which consequently impedes their ability to fully leverage AI technologies. Specifically, 67% of respondents indicated a lack of complete trust in their organization's data for decision-making. The survey underscores the imperative for further innovation and improvement in data management practices to overcome these obstacles and achieve strategic objectives.

Furthermore, a shortage of specialized skills and resources is impacting data quality and AI initiatives, identified as a prominent challenge to the success of data programs by 42% of respondents, an increase from 37% in the previous year. This skills gap is further emphasized by the finding that 60% of organizations perceive themselves as unprepared for AI initiatives in terms of employee skills and training.

The methodology involved surveying 565 data and analytics professionals worldwide in the first half of 2024, with analysis led by Drexel LeBow in collaboration with Precisely. The respondents represented a diverse range of functional titles and industries, ensuring a comprehensive perspective on the challenges and priorities in the field of data integrity and AI readiness.

Summary of Key Findings

This paper addresses the evolving landscape of workforce development in light of advancements in artificial intelligence (AI), focusing on the concept of "intelligence augmentation," where humans and AI collaborate synergistically. The central argument posits that while AI excels at calculation, computation, and prediction ("reckoning"), human judgment skills, encompassing decision-making under uncertainty, deliberation, ethics, and practical knowing, will become increasingly crucial.

The paper defines intelligence as "the disposition to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience," emphasizing a performance-based and dispositional view aligned with current research on intelligent behavior. Within this framework, intelligence is divided into two complementary roles: reckoning (AI's strength) and judgment (human's strength).

The core finding is that workforce development must shift its focus from primarily reckoning skills to cultivating human judgment skills to prepare individuals for collaborative partnerships with AI. This shift is driven by the rapid pace of AI-led changes in the division of labor between humans and machines.

The paper uses hypothetical cases involving small business owners in food services and green energy to illustrate how AI-based reckoning and human judgment can complement each other. These cases highlight the need for individuals to develop new skills to maintain economic inclusion in a future where AI plays a more prominent role. The analysis builds upon foundational work by Pearson and NESTA, which forecasted shifts in occupational skills.

Conclusion: Summary of Key Findings

This research investigated the application of offline reinforcement learning (RL) to optimize human-AI decision-making, focusing on human-centric objectives beyond mere decision accuracy. The core finding is that offline RL presents a promising approach for modeling the dynamics of human-AI interaction, leading to policies that can optimize human-centric objectives and provide insights into the AI-assisted decision-making space.

The study instantiated this approach by optimizing two specific objectives: human-AI accuracy on the decision-making task and human skill improvement (learning about the task). Policies were learned from existing human-AI interaction data and compared against several baselines.

Key findings from two experiments (N = 316 and N = 964) consistently demonstrated that participants interacting with policies optimized for accuracy achieved significantly better accuracy, including human-AI complementarity, compared to those interacting with other types of AI support. However, optimizing for human learning proved more challenging, with significant learning improvement observed only intermittently among participants interacting with learning-optimized policies.

The research methodology involved framing human-AI decision-making as a Markov Decision Process, enabling the learning of policies that adapt to context and objectives using offline RL. The reward function was designed to capture objectives like human learning or task enjoyment, while the state space incorporated human-centric and contextual factors (e.g., human skill, cognitive load, AI uncertainty). The action space comprised various human-AI interaction techniques tailored for specific contexts.

In summary, the study highlights the importance of considering human-centric objectives beyond decision accuracy in AI-assisted decision-making and establishes offline RL as a viable method for optimizing human-AI interaction for such objectives. This work opens up the research challenge of optimizing human-AI interaction for a broader range of human-centric outcomes.

Conclusion: Summary of Key Findings

This research article investigated the impact of Artificial Intelligence (AI) on decision-making processes within organizations, focusing on its consequences for individuals, companies, and society. The study highlighted AI's role in analyzing data insights, identifying gaps, and transforming decision-making proficiency, especially under time constraints. AI's capabilities in predicting business trends, transforming strategies, and enabling quick decisions were identified as key drivers of productivity and business success.

The research examined how AI transforms internal operations across departments, including transport and consultation management, by predicting demands, adjusting supply levels, and optimizing results, leading to reduced operating costs. AI software facilitates data-driven choices, improves customer targeting, enhances development, and customizes data reporting for strategic decisions. The study provided an overview of AI mechanisms such as automation efficiency, complex management, succession planning, and risk assessment across various departments (e.g., Human Resources, Accounts & Finance, Sales & Marketing) and industries (e.g., Healthcare, Finance, Consulting, Transportation).

The literature review indicated that AI-driven approaches have substantially enhanced the speed of improvement in business activities. Machine learning and natural language processing modernize and simplify processes, resulting in faster results and increased responsiveness in decision-making. AI's role in exploring client response, understanding preferences, and adapting alterations was also emphasized. The combination of AI into business processes has emerged as a key driver in increasing quickness, accuracy, and stability in data analysis.

However, the study also addressed challenges associated with AI in decision-making, including ethical concerns, biased algorithms, and social allegations. Data privacy protection, transparency, accountability, and explainability issues were identified as critical factors for a company's reputation. The importance of ethical AI methods, transparency, and nonbiased algorithms was underscored. The research emphasized the need for governance regulations and policies to mitigate biases and ensure AI systems align with organizational aims and goals. The quality of input data was identified as a critical factor, with inadequate, incorrect, or influenced data leading to adverse effects.

Conclusion: Summary of Key Findings

This research investigated the phenomenon of overreliance on AI in AI-assisted decision-making and explored the efficacy of cognitive forcing functions as a means to mitigate this issue. The study (N=199) compared three cognitive forcing designs with two explainable AI (XAI) approaches and a no-AI baseline. The core finding is that cognitive forcing functions significantly reduced overreliance on AI compared to simple XAI approaches. This suggests that interventions designed to promote analytical engagement with AI-generated explanations can be effective in improving human decision-making in AI-assisted contexts.

However, the study also revealed a trade-off between the effectiveness of cognitive forcing functions and their perceived usability. Designs that most effectively reduced overreliance received less favorable subjective ratings, indicating a potential challenge in balancing cognitive engagement with user experience. Furthermore, the research identified that cognitive forcing interventions disproportionately benefited individuals with a higher Need for Cognition (NFC), suggesting that the effectiveness of XAI solutions is moderated by human cognitive motivation.

In summary, the study introduced cognitive forcing functions as a viable interaction design intervention for reducing overreliance on AI. The findings highlight the importance of considering both the effectiveness and acceptability of such interventions, as well as the role of individual differences in cognitive motivation, when designing AI-assisted decision-making systems.

Conclusion: Summary of Key Findings

This research investigated the influence of AI agents on human decision-making within AI-augmented systems, specifically focusing on the high-uncertainty environment of Initial Coin Offering (ICO) evaluations. Data was collected from a leading ICO evaluation platform where an

AI agent initially rated ICOs based on observable and quantitative properties, followed by human expert evaluations. The study sought to determine if human decisions are influenced by AI agents and under what circumstances humans can mitigate AI-induced errors.

The analysis revealed that human experts are generally influenced by AI agents. Specifically, when the AI agent assigned a low rating to an ICO project, human experts were more likely to align with this assessment. Conversely, a high AI rating did not lead to blind adherence from human experts, suggesting a more critical evaluation in such cases. This behavior indicates that human experts utilize the AI agent as a filter, readily dismissing projects with low AI ratings while scrutinizing those with high ratings more thoroughly.

The findings suggest a combination of algorithm appreciation and algorithm aversion, where the AI agent exerts an asymmetric anchoring effect on human decision-makers. The AI agent serves as a filter, prompting closer examination of riskier (or rarer) recommendations. Furthermore, the study demonstrated that the hybrid human expert and AI system, overall, outperformed the AI agent alone. This highlights the potential for synergistic benefits when combining AI capabilities with human judgment in complex evaluation tasks.

Summary of Key Findings

This research explores an alternative approach to improving decision-making by leveraging machine learning (ML) to facilitate stakeholder deliberation on organizational decision-making processes, rather than focusing solely on automating or assisting individual decisions. The study introduces "Deliberating with AI," a web tool designed to enable stakeholders to create and evaluate ML models, thereby examining the strengths and shortcomings of past decision-making practices and deliberating on improvements for future decisions.

The methodology involves a case study in the context of people selection, specifically graduate school admissions, where both decision-makers (faculty) and decision subjects (students) utilized the tool. The key finding is that the stakeholders effectively used the ML models generated by the web tool as boundary objects to deliberate over organizational decision-making practices. This deliberation allowed for the sharing of perspectives and nuanced discussions regarding admissions decision-making, ultimately informing future decision-making practices.

The study contributes to the literature on human-centered AI and organizational decision-making by demonstrating how participatory AI design and ML-driven deliberation can be integrated to improve human decision-making. The research provides insights into stakeholder-centered participatory AI design and technology for organizational decision-making, highlighting the potential of ML models to externalize patterns, including biases, and facilitate reflection and deliberation among stakeholders. The web tool serves as a method to guide users in participatory AI design while deliberating on how to improve future personal and organizational decision-making.

Conclusion: Summary of Key Findings

This research addresses the gap between model-centric explainable AI and the need for human-centered explanations in AI-assisted decision making. The study employed two primary approaches: an empirical experiment to investigate user responses to different explanation types and their impact on trust calibration, and the development of an educational resource for industry designers to facilitate the creation of effective explanation interfaces.

The empirical experiment revealed that explanations do not consistently improve trust calibration and, in some cases, can negatively impact it, particularly among novice users with low self-competence. This finding challenges the assumption that providing explanations automatically leads to better-informed trust decisions.

Exploratory interviews and usability testing with industry designers highlighted a demand for a comprehensive and accessible educational resource. This resource should translate research findings, such as those from the experiment, into practical guidance for designing explainable interfaces for AI products. The need for such a resource underscores the importance of bridging the gap between theoretical research and practical application in the field of explainable AI.

Recommendations for Responsible AI Implementation

Conclusion: Recommendations for Responsible AI Implementation

The integration of Artificial Intelligence (AI) into higher education necessitates a framework for responsible implementation, focusing on academic integrity, data privacy, and critical AI literacy. This section outlines key recommendations derived from an analysis of AI's impact on education and its potential for misuse.

Academic Integrity and AI Usage: To uphold academic integrity, institutions should normalize AI generative text citation practices, akin to standard attribution methods for external sources. Current writing style manuals (APA, MLA, Chicago) offer guidance on citing AI-generated materials. Furthermore, instructors should clearly articulate expectations regarding AI use in syllabi and assignments, specifying acceptable parameters and requiring students to demonstrate and delineate the role of AI in their work. This transparency ensures that AI serves as a tool for learning and critical analysis rather than a shortcut that bypasses the educational process.

Data Privacy and Security: Given that most AI programs are hosted by commercial entities, stringent data privacy protocols are crucial. Institutions must emphasize the importance of protecting sensitive information, such as student data regulated by FERPA, human subject research data, health information, and confidential work-related materials. Students and faculty should treat unsecured AI systems as public platforms, refraining from sharing information that would not be publicly disclosed.

AI and Prompt Literacy: Cultivating AI literacy is paramount. Students should critically assess the ethical implications of AI, creatively apply AI tools within their fields of study, and demonstrate a commitment to responsible use. This includes understanding the strengths and weaknesses of various AI platforms and employing effective prompting techniques, such as the RACE format (Role, Action, Context, Execute), to elicit relevant and accurate responses. The development of critical thinking skills remains essential, as these skills are foundational for effective AI utilization.

Addressing Bias and Hallucinations: Recognizing the inherent biases within AI algorithms and the potential for "hallucinations" (presenting incorrect information as fact) is critical. Institutions, particularly those serving diverse populations, must actively address the ways AI systems can reinforce harmful stereotypes and perpetuate social inequalities. Users should be encouraged to provide feedback and report instances of bias or discrimination. Furthermore, students should be cautioned against accepting AI-generated information as wholly factual, particularly in niche or underrepresented topics.

AI as a Tutor and Learning Tool: AI can be leveraged as a valuable tutoring and learning tool. Generative AI tools can summarize readings, offer feedback, provide editing suggestions, teach problem-solving steps, generate annotated bibliographies, assess evidence relevance, and refine theses. AI tutoring tools can adapt to individual student progress, providing tailored guidance. However, the integration of AI as a learning tool should be carefully managed to ensure it complements, rather than replaces, traditional pedagogical approaches.

Conclusion: Recommendations for Responsible AI Implementation

Based on the information presented, responsible AI implementation in research necessitates adherence to ethical principles, robust data security measures, and careful consideration of potential biases and harms. Key recommendations derived from the document include:

Ethical Considerations and Consent:

- **Respect for Individual Rights:** Implementations should align with the principles outlined in the "Blueprint for an AI Bill of Rights," ensuring protection from unsafe systems, algorithmic discrimination, and promoting equitable design.
- **Data Governance and Transparency:** Researchers must provide transparency regarding the use of automated systems, explaining their contribution to outcomes and offering opt-out options where appropriate.
- **Informed Consent:** Consent for interaction with or data use for AI tools should not be assumed, particularly given challenges to the use of personal information in building large language models (LLMs) and the potential for biased AI outcomes.

Data Security and Privacy:

- **Data Classification and Access Control:** Adhere to Northwestern University's data classification levels, ensuring appropriate security measures are in place based on data sensitivity (Public, Sensitive, Contractual/Legal Restrictions, Classified/Export Controls).
- **Secure Data Handling:** Avoid uploading research data into unapproved tools, services, or websites. Utilize university-approved resources for human research data.
- **Third-Party Risk Assessment:** Conduct Service Provider Security Assessments through IT for all third-party tools used with Personally Identifiable Information (PII) or Protected Health Information (PHI).
- **Data Minimization and De-identification:** Recognize that removing names alone is insufficient for de-identification. Implement robust de-identification strategies to prevent re-identification through text, images, or combination with other data sources.

Bias Mitigation and Risk Management:

- **Bias Assessment and Mitigation:** Develop and document approaches to assess and mitigate bias in study protocols, particularly for new models and models used in decision-making processes. Consider independent review of bias assessments.
- **Risk Evaluation:** Evaluate potential risks of harm from re-identification, inference of private data, inaccurate output, exposure to AI-generated content, and biased models. Implement safeguards to mitigate these risks.
- **Model Evaluation:** Assess bias in outcomes for specific use cases, especially for models used in decision-making processes. Recognize that assessment checklists may be insufficient to detect all forms of bias.

Infrastructure and Resources:

- **Utilize University-Approved Resources:** Leverage platforms like Microsoft Azure OpenAI service with security controls, or local models on secure, university-owned devices.

- **Consult with Experts:** Engage with data scientists and research analysts experienced in working with AI models and tools. Contact Northwestern IT Research Computing and Data Services or Feinberg School of Medicine (FSM) IT for guidance.
- **Secure Coding Practices:** Employ secure coding practices, such as using separate chat windows for code assistance, utilizing GitHub Enterprise with appropriate security controls, and using synthetic data files.

IRB Considerations:

- **Regulatory Oversight:** Recognize that the development of AI/ML tools for diagnostic purposes should have the same degree of regulatory oversight as the development of in vitro diagnostic devices.

Conclusion: Recommendations for Responsible AI Implementation

Based on the Los Angeles County Office of Education's (LACOE) Task Force findings, responsible AI implementation in TK-12 education necessitates a multi-faceted approach encompassing ethical considerations, equity, and robust evaluation frameworks. The Task Force, comprised of diverse stakeholders including educators, administrators, students, and community partners, emphasizes the following recommendations:

1. Prioritizing Equity and Addressing Bias:

Central to responsible AI implementation is the promotion of equity. The Task Force identifies the critical need to address potential biases embedded within AI technologies to prevent the widening of educational gaps. This requires a proactive approach to ensure fair distribution of resources and equitable access to AI-powered learning opportunities for all students, particularly those from marginalized groups. Strategies include:

- Designing AI implementation to provide scaffolding and differentiation assistance for neurodivergent students, English Language Learners, and other groups to address developmental bias.
- Promoting digital citizenship and discussions to foster critical thinking about AI's potential impact on marginalized groups.
- Advocating for inclusive content creation processes and empowering marginalized groups in AI development.

2. Establishing Clear Guidelines and Evaluation Frameworks:

The Task Force advocates for the development of clear guidelines for evaluating, piloting, and using AI tools. This includes establishing decision-making processes involving diverse perspectives to prevent bias and ensuring that AI serves as a tool to support educators and promote fairness in educational opportunities. Key components of this framework include:

- **Needs Assessment:** Conducting thorough needs assessments, as exemplified by LACOE's partnership with Project Tomorrow and ASU Learning Transformation Studios, to gauge familiarity with AI, attitudes toward its use in education, and identify concerns and challenges. The survey data collected informed the development of key points, areas of need, and topics for professional learning.
- **Risk Mitigation:** Developing strategies to address potential risks associated with AI adoption, such as increased dependency on technology, data privacy concerns, and the potential for misuse.

- **Vendor Vetting:** Establishing contracting protocols to hold vendors accountable and compliant with legal requirements, industry standards, and best practices regarding bias, accuracy, security, privacy, and transparency.

3. Empowering Educators and Fostering AI Literacy:

Recognizing the transformative potential of AI, the Task Force emphasizes the importance of empowering educators with the knowledge and skills necessary to effectively integrate AI into their teaching practices. This includes:

- Providing professional learning opportunities for educators and school leaders regarding the management of AI learning environments and the appropriate use of AI tools to support teaching and learning.
- Promoting AI literacy among educators and students, enabling them to critically evaluate AI systems and tools and understand their ethical implications.
- Training users on the detection and prevention of AI misuse, equipping them with the knowledge to identify suspicious activity.

4. Ensuring Data Privacy and Security:

Protecting student data privacy and ensuring the security of learning systems are paramount. The Task Force recommends:

- Providing information about how student data is protected within AI tools and guidance toward which tools to use within the classroom.
- Understanding state and federal regulations and policies around AI usages in K-12 schools.
- Implementing robust cybersecurity measures to prevent intentional malicious use, cyber attacks, and manipulation of learning systems.

5. Continuous Monitoring and Feedback Loops:

The Task Force underscores the importance of continuous monitoring and feedback loops to ensure that AI implementation remains aligned with educational goals and ethical principles. This involves:

- Establishing a safe learning environment for students and teachers to feel comfortable sharing candid thoughts about AI implementation.
- Prioritizing student privacy, safety, and equity in all AI applications.
- Promoting responsible use of AI tools, including educating users on ethical guidelines.

Conclusion: Recommendations to Guide UC’s AI Strategy

This section synthesizes recommendations derived from the University of California Presidential Working Group on AI, focusing on responsible AI implementation within the UC system. The Working Group, composed of faculty and staff from all ten UC campuses and representatives from various UC offices, conducted research and analysis to operationalize the UC Responsible AI Principles. The methodology involved establishing four subcommittees, each dedicated to a high-risk AI application area: health, human resources (HR), policing, and student experience. Each subcommittee investigated AI use cases, assessed benefits and risks, and formulated guidance for mitigating harms and maximizing benefits. The Working Group's efforts culminated in recommendations designed to guide UC President Michael Drake in developing mechanisms to operationalize the UC Responsible AI Principles, establishing data stewardship standards, and creating coordinated campus-level councils and systemwide coordination through the University of California Office of the President (UCOP).

Conclusion: Recommendations for Responsible AI Implementation

The operationalization of AI principles represents a critical juncture for stakeholders across sectors, influencing the safe and responsible development and deployment of AI technologies globally. Early efforts in this domain are particularly influential due to technological, organizational, and regulatory lock-in effects. This paper analyzes recent initiatives aimed at translating AI principles into practice, offering valuable lessons for stakeholders seeking to develop trustworthy AI technologies.

The case study of Microsoft's AI, Ethics and Effects in Engineering and Research (AETHER) Committee demonstrates a model for integrating AI principles into corporate practices and policies. Key drivers of success include executive and employee buy-in, integration into a broader company culture of responsible AI, and the creation of interdisciplinary working groups to address issues such as bias, fairness, reliability, safety, and potential threats to human rights.

OpenAI's staged release of GPT-2 provides insights into responsible publication norms for advanced AI technologies. The company's approach highlights the importance of threat modeling, documentation schemes to promote transparency about potential risks, and open communication with users to identify and mitigate harms. This methodology offers a framework for institutions to manage the societal and policy implications of powerful AI models.

The OECD AI Policy Observatory exemplifies an intergovernmental initiative that facilitates international coordination in implementing AI principles. The Observatory's success is attributed to sustained government support, the convening of multistakeholder expert groups, and significant outreach efforts to global partners. This model provides a counterpoint to "AI nationalism" by fostering collaboration and shared understanding in AI governance.

These case studies collectively underscore the importance of proactive measures, multi-stakeholder engagement, and continuous evaluation in ensuring the responsible implementation of AI principles. The lessons learned from these initiatives can inform the development of effective strategies for navigating the complex challenges associated with AI governance.

Recommendations for Responsible AI Implementation

This research introduces the Responsible AI for Service Excellence (RAISE) framework, defined as a strategic approach for integrating AI into service industries responsibly. RAISE emphasizes collaborative AI design and deployment, aligning with evolving global standards and societal well-being while promoting business success and sustainable development. The framework is underpinned by three core principles: (1) Embrace AI to serve the greater good, (2) Design and deploy responsible AI, adhering to ethical principles of privacy, transparency, accountability, justice, fairness, and non-maleficence, and (3) Practice transformative collaboration with diverse service entities to implement responsible AI.

The RAISE framework is informed by the United Nations’ Sustainable Development Goals (SDGs) and established AI ethics guidelines. It addresses the urgent need for responsible AI deployment and usage, particularly given the rapid development and spread of generative AI (GenAI) technologies. The framework acknowledges potential risks and challenges associated with AI implementation and provides practical recommendations for service entities (service organizations, policymakers, AI developers, customers, and researchers) to strengthen their commitment to responsible and sustainable service practices.

The research advocates for an integrated approach where AI technologies are designed and implemented to contribute to business success while simultaneously serving the greater good. This perspective challenges the traditional view of profit generation and social good as mutually exclusive objectives. The emphasis on transformative collaboration highlights the need for addressing responsible practices with broader societal implications.

Conclusion

Recommendations for responsible AI implementation necessitate a structured approach to framework selection and application. The analysis of over 40 existing responsible AI frameworks reveals a critical need for organizations to systematically characterize these tools based on their specific needs. A matrix approach, focusing on the user (Development/Production teams and Governance teams) and the utility dimension (AI system components, lifecycle stages, or characteristics), is proposed as a method for organizations to precisely apply frameworks and understand their utility relative to existing guidance. The research highlights that process frameworks, while flexible and adaptable, often lack specificity, making it challenging for organizations with limited experience in responsible AI implementation to effectively leverage them. The proposed matrix aims to address this gap by providing a structured way of thinking about who could benefit from a framework and how it could be used.

Conclusion: Recommendations for Responsible AI Implementation

This section outlines recommendations for university administrators to foster responsible AI implementation, focusing on strategy, governance, risk assessment, transparency, and intellectual property compliance. These recommendations are derived from a framework emphasizing ethical considerations and accountability in the application of generative AI technologies.

Coordinated Strategy and Governance: A unified, university-wide approach to AI governance and oversight is crucial. This includes incorporating responsible AI principles into procurement policies and establishing comprehensive training programs for procurement officers.

Multistakeholder Campus-Level Councils: The establishment of campus-level councils, comprising faculty, staff, and student representatives, is recommended to ensure appropriate AI procurement, development, implementation, and monitoring. Furthermore, each college/school should establish data stewardship standards, involving relevant bodies responsible for information risk management, data privacy, and security. Seeking external stakeholder opinions, including policymakers, industry experts, and the public, is also vital for ensuring responsible generative AI research practices.

Comprehensive Risk and Impact Assessment Framework: A shared framework for AI evaluation and oversight is necessary, considering criteria such as the nature, scale, duration, irreversibility, and likelihood of adverse impacts. Tiered risk levels should be utilized to enable differentiated assessments based on potential risks.

Transparency and Accountability through Public Documentation: Establishing publicly accessible databases specifically for AI-enabled technologies that pose a higher than moderate risk to individual rights is recommended. These databases should include detailed descriptions, developers' information, intended uses, limitations, potential risks, risk mitigation strategies, and details about algorithms and training data (where possible). Standardized database structures across colleges/schools will allow for cross-college insights and comparisons. Channels for feedback and contesting decisions made by AI-enabled tools should also be provided.

Enforcing Intellectual Property Compliance: University administrators should stay abreast of legal developments in intellectual property, particularly concerning generative AI. Developing clear policies and procedures to govern the use of generative AI, covering content ownership, licensing, and copyright compliance, is essential. Educating employees on intellectual property rights, including those related to AI-generated content, and monitoring the implementation of generative AI within the university to ensure compliance with intellectual property laws are also critical. Comprehensive training for scholars and students on the effective and ethical use of generative AI tools, covering both technical aspects and ethical implications, should be provided through workshops and seminars.

Conclusion: Recommendations for Responsible AI Implementation

The Department of Defense (DOD) has articulated five ethical principles for AI systems: responsible, equitable, traceable, reliable, and governable. However, these principles require further refinement for effective implementation across the AI lifecycle, from development to operations. A foundational step is the standardization of language and definitions related to responsible AI characteristics, crucial for policymakers, engineers, program managers, and operators.

This study advocates for the DOD to adopt or adapt the National Institute for Standards and Technology's (NIST) draft taxonomy as the standardized language for responsible AI. NIST's AI Risk Management Framework (AI RMF) provides a community-defined taxonomy that the DOD can leverage. The alignment between NIST's trustworthy characteristics and the DOD's ethical principles offers a basis for collaborative development and deployment of responsible AI.

Specifically, the research recommends the DOD:

1. **Adopt terms and definitions from NIST:** Including accuracy, robustness, transparency, explainability, interpretability, and privacy-enhanced.
2. **Adapt DOD terms and definitions to NIST's terms and definitions:** Transforming "Reliable" to "Reliability," "Governable" to "Safe," and "Equitable" to "Managing Bias."
3. **Retain DOD-unique terms and definitions:** Specifically, "Responsible" and "Traceable," due to their unique relevance to the DOD's context and mission.

Adopting or adapting NIST's taxonomy offers several benefits: reducing misunderstandings among stakeholders, providing precise terms for guiding AI development and operationalization, and promoting a consistent U.S. government stance in international discussions on responsible AI norms. This approach facilitates clear communication, reduces friction, and supports the development and deployment of responsible AI systems within the DOD.

References

- [Author, A. A.]. (Year). [Title of paper: Background and significance of AI in decision making]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Background and Significance of AI in Decision Making] [inferred]. Retrieved from 9.pdf [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Background and Significance of AI in Decision Making] [inferred]. Retrieved from 8.pdf [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Background and Significance of AI in Decision Making] [inferred]. Retrieved from 7.pdf [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Background and Significance of AI in Decision Making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Background and significance of AI in decision making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper]. [inferred] [inferred]

- [Author, A. A.]. (Year). [Title of paper: Background and significance of AI in decision making]. [inferred].

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper]. [inferred]. *[Journal Name, inferred]*, *[Volume, inferred]*([Issue, inferred]), [Pages, inferred]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Background and Significance of AI in Decision Making] [inferred].

- [Author, A. A.]. (Year). [Title of paper]. [inferred]

- [Author, A. A.]. (2019). [Title of paper: Thesis Statement and Research Questions]. [inferred]

- [Author, A. A.]. (Year [inferred]). [Title of paper: Thesis Statement and Research Questions] [inferred].

- [Author, A. A.]. (Year [inferred]). [Title of paper: Thesis Statement and Research Questions] [inferred].

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Thesis Statement and Research Questions]. [inferred].

- [Author, A. A.]. (Year). [Title of paper on Thesis Statement and Research Questions]. [inferred]

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Thesis Statement and Research Questions]. [inferred]. Retrieved from 14.pdf

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Thesis Statement and Research Questions]. [inferred].

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Thesis Statement and Research Questions]. [inferred].

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Thesis Statement and Research Questions]. [inferred]. Retrieved from 11.pdf

- [Author, A. A.]. [Title of paper: Theoretical Frameworks of Human Decision Making]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Theoretical Frameworks of Human Decision Making]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Theoretical Frameworks of Human Decision Making]. [inferred]
- [Author, A. A.]. [Title of paper: Theoretical Frameworks of Human Decision Making]. [inferred].
- [Author, A. A.]. (n.d.). [Title of paper based on Theoretical Frameworks of Human Decision Making]. [inferred]
- [Author, A. A.]. (Year). [Title of paper: Theoretical frameworks of human decision making]. [inferred].
- [Author, A. A.]. (n.d.). [Title of paper: Theoretical Frameworks of Human Decision Making] [inferred]. Retrieved from 23.pdf [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: Theoretical Frameworks of Human Decision Making] [inferred]. Retrieved from 22.pdf [inferred].
- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Theoretical frameworks of human decision making]. [inferred]. Retrieved from 21.pdf
- [Author, A. A.]. (Year). [Title of paper: Theoretical frameworks of human decision making]. [inferred].
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes]. [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes]. [inferred].
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes]. [inferred].
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes]. [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes]. [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes] [inferred]. [inferred from 34.pdf].
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes]. [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes] [inferred]. [inferred from 32.pdf].
- [Author, A. A.]. (n.d.). [Title of paper: AI's influence on cognitive processes] [inferred].
- [Author, A. A.]. (Year). [Title of paper: AI's influence on cognitive processes]. [inferred].
- [Author, A. A.]. (2024). [Title of paper: Ethical considerations of AI-driven decisions]. [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: Ethical considerations of AI-driven decisions] [inferred]. Retrieved from 47.pdf [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: Ethical considerations of AI-driven decisions]. [inferred].

- [Author, A. A.]. (Year, inferred). [Title of paper: Ethical considerations of AI-driven decisions] [inferred]. Retrieved from 45.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Ethical considerations of AI-driven decisions] [inferred]. Retrieved from 44.pdf [inferred]

- [Author, A. A.]. (Year). [Title of paper: Ethical considerations of AI-driven decisions] [inferred]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Ethical considerations of AI-driven decisions]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Ethical considerations of AI-driven decisions]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Ethical considerations of AI-driven decisions] [inferred]. Retrieved from 40.pdf [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Ethical considerations of AI-driven decisions] [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Research design and data collection methods]. [inferred].

- [Author, A. A.]. [inferred] ([Publication Year, inferred]). [Title of paper: Research design and data collection methods]. [inferred]. Retrieved from 55.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Research design and data collection methods]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Research design and data collection methods]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Research design and data collection methods]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Research design and data collection methods]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Research Design and Data Collection Methods]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Research design and data collection methods]. [inferred].

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Research design and data collection methods]. [inferred]. Retrieved from 48.pdf

- [Author, A. A.]. (n.d.). [Title of paper on Research Design and Data Collection Methods]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Participants and sampling techniques]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Participants and sampling techniques]. [inferred]

- [Author, A. A.]. [inferred] ([Publication Year, inferred]). [Title of paper: Participants and sampling techniques]. [inferred]. Retrieved from 63.pdf

- [Author, A. A.]. [inferred] ([Publication Year, inferred]). [Title of paper: Methodology - Participants and Sampling Techniques]. [inferred]. Retrieved from 62.pdf

- [Author, A. A.]. [inferred] ([Inferred Year]). [Title of paper: Methodology - Participants and Sampling Techniques]. [inferred]. Retrieved from local file 61.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Participants and sampling techniques]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Participants and sampling techniques]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Participants and sampling techniques]. [inferred]

- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Methodology - Participants and Sampling Techniques]. [inferred]. Retrieved from 57.pdf

- [Author, A. A.]. [inferred] ([Publication Year, inferred]). [Title of paper: Methodology - Participants and Sampling Techniques]. [inferred]. Retrieved from 56.pdf

- [Author, A. A.]. [inferred] ([Publication Year, inferred]). [Title of paper: Data analysis techniques]. [inferred]. Retrieved from local file 75.pdf

- [Author, A. A.]. [inferred] ([Publication Year inferred from filename]). [Title of paper: Data analysis techniques]. [inferred]. Retrieved from 74.pdf

- [Author, A. A.]. [inferred] ([Publication Year inferred from filename]). [Title of paper: Data analysis techniques]. [inferred]. Retrieved from 73.pdf

- [Author, A. A.]. (Year inferred from filename). [Title of paper: Data analysis techniques]. [inferred]. Retrieved from 72.pdf [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Data analysis techniques methodology]. [inferred]. Retrieved from 71.pdf [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Data analysis techniques methodology]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Data analysis techniques methodology]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Data analysis techniques methodology]. [inferred].

- [Author, A. A.]. (2006). [Title of paper: Data analysis techniques]. [inferred]. Retrieved from 67.pdf [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Data analysis techniques methodology]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Positive impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Positive impacts of AI on decision making] (Report No. 84). [inferred]

- [Author, A. A.]. (Year). [Title of paper: Positive impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Positive impacts of AI on decision making]. [inferred].

- [Author, A. A.]. (Year, inferred). [Title of paper: Positive impacts of AI on decision making] [inferred]. Retrieved from 81.pdf

- [Author, A. A.]. (Year inferred). [Title of paper: Positive impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Positive impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Positive impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Positive impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Positive impacts of AI on decision making] (76.pdf). [Inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Negative impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Negative impacts of AI on decision making] [inferred]. Retrieved from 93.pdf [inferred]

- [Author, A. A.]. (Year). [Title of paper: Negative impacts of AI on decision making]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Negative impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (Year inferred from filename). [Title of paper: Negative impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Negative impacts of AI on decision making]. [inferred].

- [Author, A. A.]. (Year [inferred]). [Title of paper: Negative impacts of AI on decision making] [inferred]. Retrieved from 88.pdf

- [Author, A. A.]. (Year inferred from 87.pdf). [Title of paper: Negative impacts of AI on decision making]. [inferred]

- [Author, A. A.]. (Year). [Title of paper: Negative impacts of AI on decision making]. [inferred].

- [Author, A. A.]. (Year inferred from arXiv ID). [Title of paper: Negative impacts of AI on decision making]. [inferred].

- [Author, A. A.]. (Year). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 103.pdf [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 102.pdf [inferred].

- [Author, A. A.]. (Year [inferred]). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 101.pdf

- [Author, A. A.]. (Year [inferred]). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 100.pdf

- [Author, A. A.]. (Year [inferred]). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 99.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Impact on specific decision-making contexts]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Impact on specific decision-making contexts]. [inferred].

- [Author, A. A.]. (Year [inferred]). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 96.pdf

- [Author, A. A.]. (Year [inferred]). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 95.pdf

- [Author, A. A.]. (Year [inferred]). [Title of paper: Impact on specific decision-making contexts]. [inferred]. Retrieved from 94.pdf

- [Author, A. A.]. [inferred] ([Publication Year, inferred]). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred]. Retrieved from 112.pdf

- [Author, A. A.]. (YYYY). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred].

- [Author, A. A.]. (Year). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred]

- [Author, A. A.]. (Year [inferred]). [Title of paper: Interpretation of findings and comparison with existing literature] [inferred]. Retrieved from 109.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred].

- [Author, A. A.]. [inferred] ([Publication Year, inferred]). [Title of paper: Interpretation of Findings and Comparison with Existing Literature]. [inferred]. Retrieved from 107.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred]

- [Author, A. A.]. (n.d.). [Title of paper: Interpretation of findings and comparison with existing literature]. [inferred].

- [Author, A. A.]. (Year inferred from 121.pdf). [Title of paper: Limitations of the study]. [inferred].

- [Author, A. A.]. (Year). [Title of paper: Limitations of the study]. [inferred] [Unpublished manuscript]. [inferred] Retrieved from 120.pdf

- [Author, A. A.]. [inferred]. ([Publication Year inferred from 119.pdf]). [Title of paper: Limitations of the study]. [inferred]. Retrieved from 119.pdf

- [Author, A. A.]. (YYYY). [Title of paper: Limitations of the study]. [inferred]

- [Author, A. A.]. (Year [inferred]). [Title of paper: Limitations of the study]. [inferred]. Retrieved from 117.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Limitations of the study]. [inferred] Retrieved from 116.pdf [inferred]

- [Author, A. A.]. (Year inferred from 115.pdf). [Title of paper: Limitations of the study]. [inferred].

- [Author, A. A.]. (Year inferred from 114.pdf). [Title of paper: Limitations of the study]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Limitations of the study]. [inferred] Retrieved from 113.pdf

- [Author, A. A.]. (Year inferred from 112.pdf). [Title of paper: Limitations of the study]. [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Implications for future research and practice]. [inferred].

- [Author, A. A.]. (Year). [Title of paper: Implications for future research and practice]. [inferred]

- [Author, A. A.]. [inferred] ([Inferred Year]). [Title of paper: Implications for future research and practice]. [inferred]. Retrieved from 128.pdf

- [Author, A. A.]. (YYYY). [Title of paper: Implications for future research and practice]. [inferred]

- [Author, A. A.]. [inferred] ([Inferred Year]). [Title of paper: Implications for future research and practice]. [inferred]. Retrieved from local file 126.pdf

- [Author, A. A.]. (n.d.). [Title of paper: Implications for future research and practice]. [inferred]

- [Author, A. A.]. (Year [inferred]). [Title of paper: Implications for future research and practice]. [inferred]. Retrieved from 124.pdf
- [Author, A. A.]. (n.d.). [Title of paper: Implications for future research and practice]. [inferred]. Retrieved from 123.pdf [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: Implications for future research and practice]. [inferred]
- [Author, A. A.]. [inferred] ([Inferred Year based on filename]). [Title of paper related to discussion implications for future research and practice]. [inferred]. Retrieved from 121.pdf
- [Author, A. A.]. (Year). [Title of paper: Summary of key findings on conclusion]. [inferred].
- [Author, A. A.]. (YYYY). [Title of paper: Summary of key findings on conclusion]. [inferred]. Retrieved from 139.pdf [inferred].
- [Author, A. A.]. (YYYY). [Title of paper: Summary of key findings on conclusion]. [inferred]
- [Author, A. A.]. (YYYY). [Title of paper: Summary of key findings on conclusion]. [inferred]
- [Author, A. A.]. [inferred]. ([Publication Year, inferred]). [Title of paper: Summary of key findings related to conclusion]. [inferred]. Retrieved from 136.pdf
- [Author, A. A.]. (Year [inferred]). [Title of paper: Summary of key findings on conclusion]. [inferred]. Retrieved from 135.pdf
- [Author, A. A.]. [inferred]. ([Publication Year inferred from 134.pdf]). [Title of paper: Summary of key findings on conclusion]. [inferred]. Retrieved from 134.pdf
- [Author, A. A.]. (Year inferred from 133.pdf). [Title of paper: Summary of key findings on conclusion]. [inferred].
- [Author, A. A.]. (YYYY). [Title of paper: Summary of key findings on conclusion]. [inferred]
- [Author, A. A.]. (YYYY). [Title of paper: Summary of key findings on conclusion]. [inferred]. Retrieved from 131.pdf [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: Recommendations for responsible AI implementation]. [inferred]
- [Author, A. A.]. (n.d.). [Title of paper: Recommendations for responsible AI implementation]. [inferred].
- [Author, A. A.]. (n.d.). [Title of paper: Recommendations for responsible AI implementation]. [inferred]
- [Author, A. A.]. (2023). [Title of paper: Recommendations for responsible AI implementation]. [inferred]. Retrieved from 146.pdf [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Recommendations for responsible AI implementation] [inferred]. Retrieved from 145.pdf [inferred]

- [Author, A. A.]. (Year). [Title of paper: Recommendations for responsible AI implementation]. [inferred]. Retrieved from 144.pdf [inferred].

- [Author, A. A.]. (n.d.). [Title of paper: Recommendations for responsible AI implementation]. [inferred]. Retrieved from 143.pdf [inferred]

- [Author, A. A.]. (YYYY). [Title of paper: Recommendations for responsible AI implementation]. [inferred]. Retrieved from 142.pdf [inferred]

- [Author, A. A.]. (2024). [Title of paper: Recommendations for responsible AI implementation]. [inferred]. Retrieved from 141.pdf [inferred].