Handout Week 12: AI, Bias, and the "Good Life"

Part 1: Small Group Breakouts (15-20 Minutes)

Divide into groups of 3–4. Each group should choose one "Case" to discuss.

Case A: The "Mirror" vs. The "Ideal"

Some argue that if a search engine shows fewer women as "math professors," it's just reflecting the current world. Others argue we should program the AI to show more women to change future perceptions.

- **Question:** Should AI be a **mirror** (showing the world as it is) or a **mold** (shaping the world as it *should* be)? What are the risks of "correcting" the data? .

Case B: The "Proxy" Trap

If a bank removes "race" from its loan algorithm to be fair, the AI might instead use "postcode" or "shopping habits" to identify the same groups and deny them loans.

- **Question:** If bias is "baked into" our society's data, is a truly "unbiased" algorithm even possible?.

Case C: The Meaning of Work

AI is predicted to take over many tasks, potentially leading to mass unemployment or a "leisure society".

- **Question:** If you didn't *have* to work for a paycheck, what would give your life meaning? Would you feel "liberated" or "purposeless"?.

AI Ethics: Chapter 9 Knowledge Check

1. According to the text, why is bias in AI often described as "unintentional"?

- **A)** Most developers are intentionally trying to create exclusive systems.
- **B)** Developers may lack diversity or fail to imagine the potential unintended consequences for different stakeholders.
- **C)** Bias only occurs when an algorithm is explicitly programmed to be discriminatory.
- **D)** Legal protections ensure that all bias is filtered out before an AI is released.

2. What is a "proxy" in the context of algorithmic bias?

- **A)** A human supervisor who checks every decision the AI makes.
- **B)** A secondary variable (like a postcode) that the AI uses to identify protected characteristics like race.
- **C)** A test version of an algorithm used only in laboratory settings.
- **D)** A legal representative for individuals who have been harmed by AI.

3. The "mirror view" of training data suggests that:

- **A)** AI should prioritize images of historically marginalized groups to change societal perceptions.
- **B)** Training data should be an "idealized" version of how we want the world to look.
- **C)** AI should reflect the world exactly as it is, even if that includes existing societal prejudices.
- **D)** Developers should literally look in a mirror to understand their own personal biases.

4. Why does the author suggest we might want to "reserve work for humans" even if AI can do it?

- **A)** Because AI is currently incapable of performing complex cognitive tasks.
- **B)** Because work provides humans with social connection, purpose, and a sense of belonging.
- **C)** Because machines are too expensive to maintain compared to human labor.
- **D)** Because the text argues that "leisure" is actually harmful to human health.

5. What is the primary concern regarding the "speed" of ethics (the Owl of Minerva problem)?

- **A)** Ethical discussions happen so fast that technology cannot keep up.
- **B)** AI is learning to be ethical faster than humans can teach it.
- **C)** Philosophical reflection often arrives only after technology has already fundamentally changed society.
- **D)** Ethical wisdom is a universal constant that never c