Student Question

Can ethics ever be effectively encoded into AI?

AI is making more decisions that affect our lives every day—but can a machine really know what's right and wrong? Researchers say that while there are many ethical guidelines for AI, turning them into actual code is extremely tricky. Ethics depends on context, human values, and judgment—things machines don't truly understand. This raises a fascinating question: will AI ever be able to act ethically on its own, or will humans always need to make the tough moral calls?

My Response

While encoding ethics into AI is challenging due to context-dependency, Brey's concept of 'morally opaque' features suggests that encoding ethics may create new hidden ethical problems, requiring ongoing human oversight rather than full automation

You've identified a critical challenge in AI ethics: the gap between abstract ethical principles and practical implementation. Your point about context-dependency is especially important, and it connects directly to concepts we explored in our Week 1 readings.

My position is that while encoding ethics into AI is technically challenging, the process itself may create new ethical problems that require permanent human oversight rather than full automation

Brey's (2004) disclosive method is particularly relevant here. He distinguishes between "morally transparent" features that users understand and "morally opaque" features embedded in technology that have hidden moral implications. When we try to encode ethics into AI, we face a paradox: the very act of encoding may create new morally opaque features that users and even developers don't fully recognize. For example, an AI system programmed to be "fair" might operationalize fairness in ways that contradict human intuitions about justice in specific contexts.

Microsoft's "The Future Computed" (2018) acknowledges this tension, proposing six ethical principles for AI, including fairness, reliability, and accountability, but emphasizes that "determining the full range of work needed to address possible bias in AI systems will require ongoing discussions" (p. 59). This suggests that ethics in AI isn't a one-time encoding problem but an iterative, socially-negotiated process.

In cybersecurity, this becomes even more complex. Should an AI-powered intrusion detection system prioritize security (blocking potentially innocent users) or privacy (allowing more permissive access)? Moor's (1985) concept of "policy vacuums" applies here; we lack clear frameworks for these novel scenarios, and simply encoding existing ethical rules may not address the new possibilities AI creates.

Follow-up question: If ethics can't be fully encoded, should we focus instead on building AI systems that can explain their reasoning to humans who then make the ethical decisions? Or does that approach create unacceptable delays in time-sensitive situations like cybersecurity incidents?

References

Moor, J. H. (1985). *What is computer ethics? Metaphilosophy, 16*(4), 266–275. https://doi.org/10.1111/j.1467-9973.1985.tb00173.x

Brey, P. (2000). *Disclosive computer ethics*. *ACM SIGCAS Computers and Society, 30*(4), 10–16. https://doi.org/10.1145/572260.572264

Microsoft Corporation. (2018). *The future computed: Artificial intelligence and its role in society* (B. Smith & H. Shum, Forewords).
https://news.microsoft.com/cloudforgood/_media/downloads/the-future-computed-english.pdf