Introduction to ethics and the course

P. J. Nickel

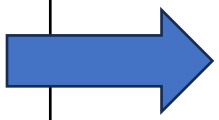
Associate Professor, Dept of Philosophy & Ethics

Introducing the course and topic

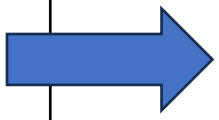
Two main topics in the course

- Ethics of AI: lectures and assignment
- Human-AI interaction: mainly in assignment

Main grade components
Written Assignment (50% of overall grade)
Final Examination (50% of overall grade)



Complete the steps as described in lectures, tutorials, and on Canvas.



Attend lectures, read the recommended texts, and study them.

Lectures and guided self-study

- In the first 5-6 weeks, there will be two lectures per week covering basic concepts applied to AI. Sometimes we'll also do activities during lecture. There will also be a Tuesday afternoon guided tutorial period where you will work on building-blocks for the assignment. I will be there to give instruction and input.
- In the first 2 weeks, you will form a group and determine a topic for the assignment.
- Your group will meet with me three times during the quarter.
- In the last couple of weeks, you will work on the main assignment and study for the exam.
- You should complete the modules each week. These are in the Modules section of the Canvas page. Some of them are time-sensitive.

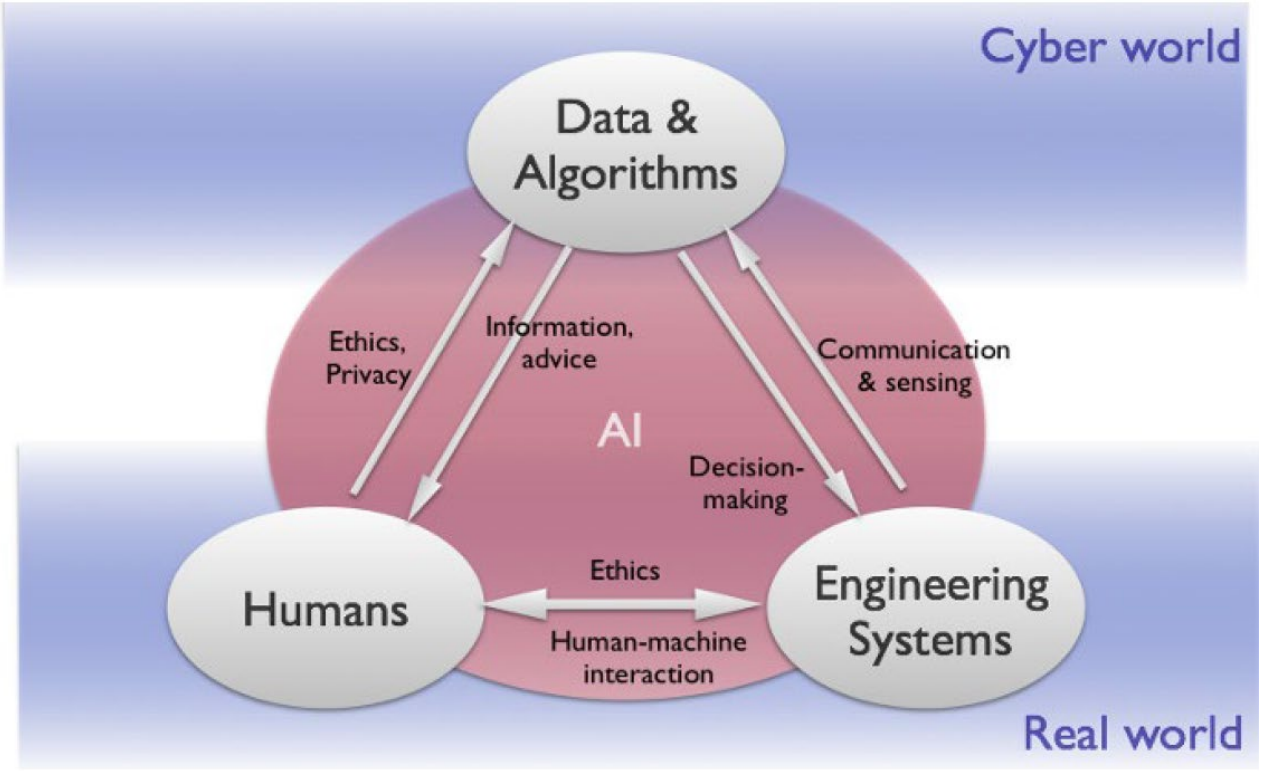
Why study ethics of AI?

- AI systems are meant to support or replace human decision-making. Ethics aims to answer the question how decisions should be made. Hence ethics can be used to guide AI design.
- Some AI designs and features have an ethical purpose (e.g., protecting privacy, achieving fair outcomes). Therefore, studying ethics will help understand and guide these designs and features.
- AI has features of speed, scale, opacity, and independence of human oversight that require special ethical attention.
- There is serious societal controversy and anxiety about AI.

How to study

- Attend the lectures and do the readings.
- Read the material twice, once for orientation, and once after the lectures to identify places where the lectures and readings overlap. The overlap forms the main source of exam questions.
- Exam questions are both open (short paragraph) and closed (MC/ T-F)
- On the exam, there are usually two ways to arrive at a correct answer. The first is to remember points directly from readings and lectures; the second is to be able to reason to the correct answer from your understanding of the material.

Relation to AI&ES program tracks



1. High-tech systems and robotics
2. Mobility
3. Healthcare
4. Smart cities
5. Science and discovery
6. Manufacturing systems

What is ethics?

- Ethics vs. morality
- Ethics is defined as the systematic study of morality. One of its tasks is to establish consistent and well-justified standards for moral judgements and reactions (e.g., assigning praise and blame).
- Law vs. ethics

Structure of ethics

We seek “reflective equilibrium” between the elements here. “Wide reflective equilibrium” refers to the need to take science and empirical factors into account, as well.

- Judgments in concrete cases (morality)
  - E.g., the “toeslagenaffaire” is blameworthy

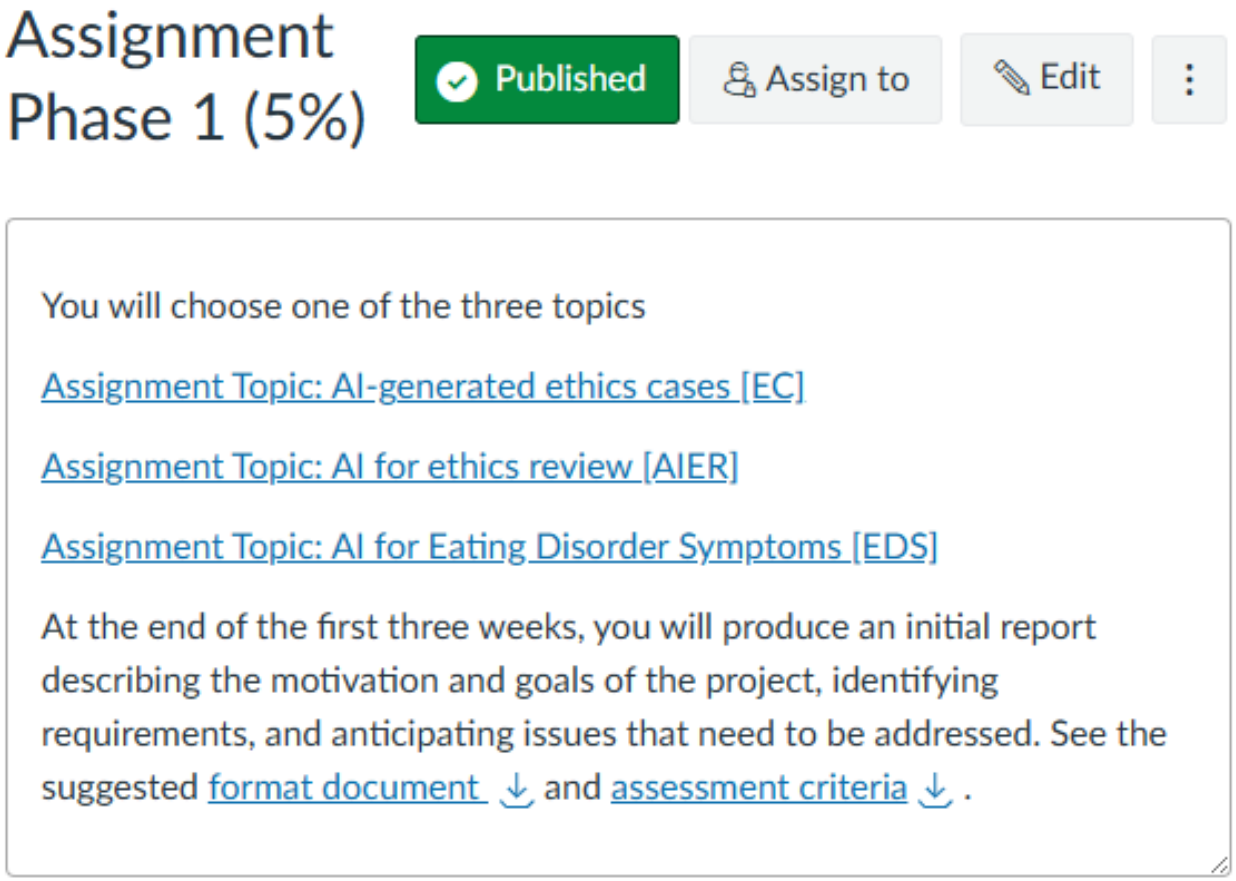
- Principles, values, virtues (ethics = systematized morality)
  - E.g., Rawls’ Difference Principle; Principle of Utility; Value of Explicability

- Theories (views about which principles, values, virtues have priority, and what justifies them)
  - Rawls’ Theory of Justice; the Capability Approach; Utilitarianism

The written assignment

[Assignment Phase 1 (5%) (tue.nl)]

Due 19 Sept



Assigned texts/ recommended readings

The primary reading is the first section of the book:

Smuha NA, ed. (2025). *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Cambridge Law Handbooks. Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-the-law-ethics-and-policy-of-artificial-intelligence/contents/B58479AF2A9C398BD461575A18F828D3

The exact schedule and readings can be accessed from the home page on Canvas.

Read with the following questions in mind:

What is the purpose of this chapter, and what is/ are the main question(s) the authors are addressing?

What are the authors' main conclusions or main claims?

What evidence or reasons do the authors supply to support the conclusions?

Do you agree with the authors' conclusions?

How might someone object to the authors' arguments?

Fundamental sources: The UDHR and the Belmont Report

Learning goals relevant to today's lecture

"Recognize and assess ethical problems in AI"

- You can **recognize some examples** as being linked with the UDHR rights and the Four Principles.

"Identify ethical issues and value conflicts at different stages of the design process."

- You are learning how to **identify ethical issues** by seeing how certain cases violate ethical rights and principles during the research and implementation phases.

By writing some thoughts down, you are also learning how to **describe** these issues and conflicts. You will continue to work on this in the written assignment.

Fundamental sources

The fundamental sources I will discuss are systematic ethical standards that have been agreed upon by human societies:

- *The Universal Declaration of Human Rights* (1948)
- *The Belmont Report* (1979)

These two sources do not mention artificial intelligence at all. But we can still *apply* them to situations in which AI is being developed or deployed.

The Universal Declaration of Human Rights (UDHR)

- Created as one of the first and most significant acts of the United Nations in 1948.
- Responds to the moral atrocities committed in World War II.
- Written collectively by delegates to the UN. The guiding force behind the process was former US first lady Eleonor Roosevelt.
- Articulates 30 rights of many different sorts.
- Ratified by many countries around the world and has a role in international law. The rights that it proclaims are therefore both moral and legal.
- Human rights are “rights possessed by all human beings (at all times and all places), simply in virtue of their humanity” (Simmons 2001).

Example

Article 12

No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

Logical relationships between rights and duties

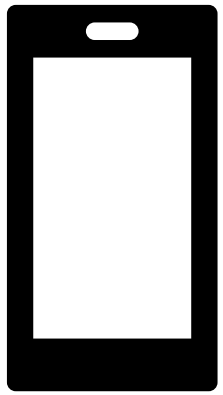
General idea: Negative or positive right ←→ Negative or positive duty

R has a right to X → S has a duty not to interfere with X (negative right)
R has a right to X → S has a duty to provide X (positive right)

Examples: family rights, rights to vote, rights to political assembly

Note that S can refer to different or multiple entities:
usually fellow citizens, institutional entities, and/or the government.

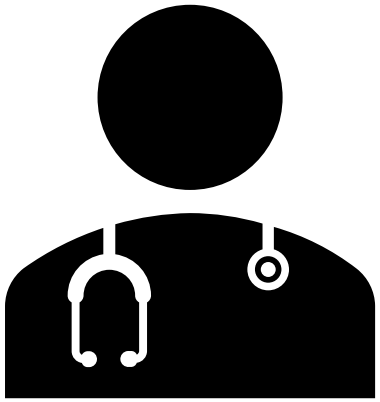
Do others have a duty not to interfere with R having X?



How strong is R’s right to X?

Do others have a duty to assist R X-ing?

Does the right to X override other rights?



Does society have a duty to help pay for R’s X-ing?

Or its consequences?

S. Matthew Liao, Human rights and AI

Fundamental Conditions Approach [FCA]: “human rights protect the fundamental conditions for pursuing a good life” (2020, 15).

Human rights have serious implications for the use of AI:

- “companies and … AI researchers are responsible for ensuring that they do not use the user’s data in ways that could undermine that user or some other users’ human rights” (ibid., 17).
- At the societal level, solutions may be needed: “one proposed solution to automation [and resulting job loss] is a universal basic income…” (ibid., 18).

S. Matthew Liao, A short introduction to the ethics of AI, *Ethics of Artificial Intelligence* (Oxford University Press, 2020): 1-42.

Consider AI risks related to the UDHR human rights

- Read and think about some of the human rights in the UDHR. You can find a link at the Modules page, or simply search.
- In your opinion, what are the human rights most likely to be *supported* by the development of AI in the next 30 years? What are the human rights most likely to be *threatened* by the development of AI in the next 30 years?
- Please describe a short example of an AI application that would threaten the right in question.

Second Fundamental Source: The Belmont Report

- The Belmont Report | HHS.gov

THE BELMONT REPORT

Office of the Secretary

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

April 18, 1979

AGENCY: Department of Health, Education, and Welfare.

ACTION: Notice of Report for Public Comment.

Differences and similarities with the UDHR

Unlike the UDHR,

- The Belmont Report has no official status in itself; it is influential because it has been adopted as the general ethical framework for human research ethics in the USA and Europe.
- The Belmont Report is framed in terms of ethics principles rather than rights.

Like the UDHR,

- The Belmont Report is multiple-authored and results from a process of consensus building.
- The Belmont Report was written as a response to serious moral failures.

“Four Principles” Approach

Beneficence

Non-maleficence

Respect

Justice

Emergence of human research ethics in the USA

http://www.tuskegee.edu/sites/www/Uploads/images/TopNews/Syphllis-survivors-with-Clinton1.jpg

Rothman, D.J. Were Tuskegee and Willowbrook 'studies in nature?' *Hastings Center Report* 12: 1982

What makes Tuskegee and related cases wrong?

The most obvious question, which has not received sufficient attention, is: how could an experiment that seems so clearly to violate medical ethics have been started and have continued for so long? It is both too simple and condescending to the researchers to argue that the primitive state of medical research ethics in the 1930s excuses the venture. As early as 1865 Claude Bernard asserted this familiar and still valid dictum: "The principle of medical and surgical morality, therefore, consists in never performing on man an experiment which might be harmful to him to any extent, even though the result might be highly advantageous to science, i.e., to the health of others."

Respect and Justice Principles were innovations meant to deal with such cases

Intuitively, there was something wrong with Tuskegee from its inception: “using people”.

The principles of beneficence and non-maleficence do not fully explain this wrong.

Hence, we need one or more additional principles to explain this wrong.

Nuremberg Code - UNC Research

Human subjects research today

In order to prevent scientific misconduct as well as to ensure that scientific research promotes the “greater good”, all human subjects research must be approved by an Ethical Review Board (ERB).

- TU/e Ethical Review Board

ERB approval is mandatory for all human subjects research, most notably **biomedical** and **psychological** research.

- But what about **AI** research?

AI as human subjects research

Artificial intelligence conducts **human subjects research** insofar as:

- The data it processes is *about* humans.
- The knowledge being acquired is used to drive decisions and develop technologies that *affect* human beings.
- Moral problems in this field are especially tricky if unequal power relationships are involved.

However, only human subjects research done with government funding, or in certain domains, is legally subject to institutional ethics review.

Facial recognition and respect for persons

“Data without consent

For facial-recognition algorithms to work well, they must be trained and tested on large data sets of images, ideally captured many times under different lighting conditions and at different angles. In the 1990s and 2000s, scientists generally got volunteers to pose for these photos — but most now collect facial images without asking permission.

For instance, in 2015, scientists at Stanford University in California published a set of 12,000 images from a webcam in a San Francisco café that had been live-streamed online[2]. The following year, researchers at Duke University in Durham, North Carolina, released more than 2 million video frames (85 minutes) of footage of students walking on the university campus[3].”

Richard van Noorden, “The ethical questions that haunt facial-recognition research”, *Nature* [News Feature], 18 November 2020. https://www.nature.com/articles/d41586-020-03187-3

In the news

NOS Nieuws • Vandaag, 09:06

Boete van privacywaakhond voor verzamelaar van miljarden foto's van gezichten

Het omstreden Amerikaanse bedrijf Clearview AI, bekend van zijn gezichtsherkenningssoftware, krijgt een boete van 30,5 miljoen euro van de Autoriteit Persoonsgegevens (AP). De privacywaakhond zegt dat Clearview een illegale database heeft opgebouwd met miljarden foto's van gezichten, ook van Nederlanders.

- Boete van privacywaakhond voor verzamelaar van miljarden foto's van gezichten (nos.nl)

What can we infer from the Four Principles about facial recognition research using public data?

- Non-maleficence
- Beneficence
- Respect
- Justice

“Respect for persons incorporates at least two ethical convictions: first, that individuals should be treated as autonomous agents, and second, that persons with diminished autonomy are entitled to protection. The principle of respect for persons thus divides into two separate moral requirements: the requirement to acknowledge autonomy and the requirement to protect those with diminished autonomy.”

“Who ought to receive the benefits of research and bear its burdens? This is a question of justice, in the sense of "fairness in distribution" or "what is deserved." An injustice occurs when some benefit to which a person is entitled is denied without good reason or when some burden is imposed unduly.”

[Belmont Report]

Facial recognition case

EU draft guidelines on facial recognition for law enforcement

“The fact that a photograph has been manifestly made public (Art. 10 LED) by the data subject does not entail that the related biometric data, which can be retrieved from the photograph by specific technical means, is considered as having been manifestly made public. Default settings of a service, e.g. making templates publicly available, or absence of choice, e.g. templates are made public without the user to be able to change this setting, should not in any way be construed as data manifestly made public.”

European Data Protection Board, “Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement” Version 1.0, Adopted on 12 May 2022. https://edpb.europa.eu/system/files/2022-05/edpb-guidelines_202205_frtlawenforcement_en_1.pdf

The EU Approach to Policy

- The following elements are part of the general framework for the EU's approach to AI policy:
  - General approach: RISK-Based
  - Legal AI (respects other EU rules and fundamental rights)
  - Robust AI
  - Ethical AI:
    - Based on "Four Principles Plus One" but expanded to include human oversight, safety, privacy, transparency, fairness, and sustainability (see CHLEP Ch. 3, p. 63).
    - The "Plus One" is *explicability* (explainability).

Questions?

Assignment topics

- You can sign up in person for a topic on Thursday 4 Sept.
- Work is done in groups of four.
- There is a limited number of slots per topic. You might not be able to work on your most preferred topic.
- The client meeting is meant to help you get ideas about what tools might be most useful, fun, or promising.

Topic 1: AI-generated ethics cases

Lay prompts can be used to generate ethics cases:

“Please compose a short description of a real-life ethical dilemma involving a conflict between confidentiality and beneficence in a clinical setting. The description should involve a patient and a GP first-line clinician. The patient should be familiar to the doctor and should have a bodily complaint.”

GPT response 1

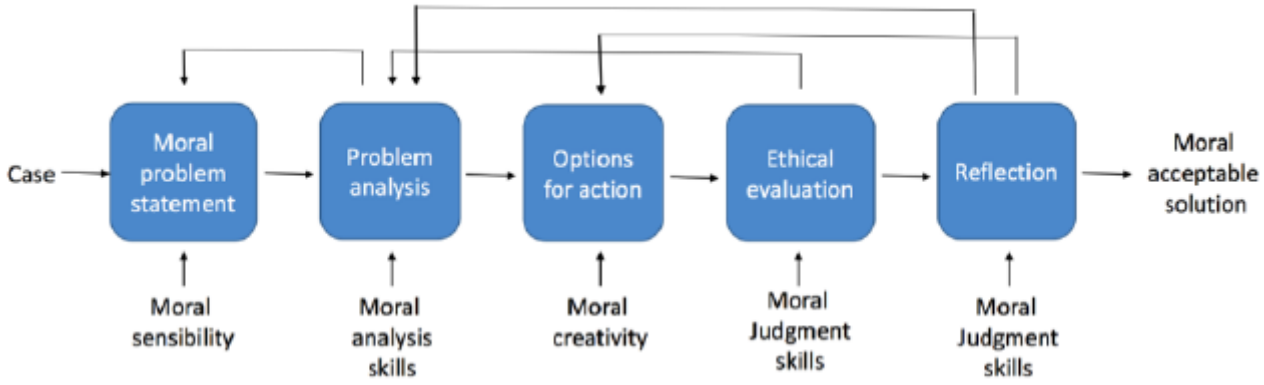
- A GP, Dr. Williams, has been seeing a patient, John, for several years. John comes in complaining of persistent abdominal pain, which, after some tests, suggests possible liver damage. During the consultation, John confides that he's been secretly consuming large amounts of alcohol, which he attributes to stress at home, specifically domestic abuse incidents involving his partner. He asks Dr. Williams not to document the alcohol consumption in his medical records, fearing that it could impact his relationship and legal situation.
- Dr. Williams faces an ethical dilemma: respecting John’s confidentiality, which is crucial to maintaining trust, versus the duty of beneficence to ensure John receives the appropriate care. Withholding this information might limit John's access to necessary support and treatments. Additionally, concealing the alcohol abuse could risk further harm to John's health, creating a conflict between respecting his wishes and acting in his best interest.

What are ethics cases?

- Cases are descriptions of a situation that has ethically salient features, usually for the purpose of eliciting moral intuitions or reactions.
- Some ethics cases are fictional (e.g., the Trolley Problem), and others are drawn from real life.

Challenges of cases in ethics education (Veraart et al. 2022)

- Are scientifically accurate
- Are interesting to students
- Help to achieve desired learning goals in the target context
- Have the right structure
- Are accessible and easy to use



Topic 2: AI supported ethics review

- The Belmont Report led to a worldwide adoption of ethics review as a standard requirement for funding and publication.
- Ethics review takes up considerable resources and causes delays: “From the academic hallways to the literature, characterizations of REBs and the research ethics review process are seldom complimentary. While numerous criticisms have been levelled, it is the time to decision that is most consistently maligned.” (Page & Nyeboer 2017).
- Because it is highly text-intensive, it is highly susceptible to partial automation using AI.
- However, there are unexplored ethical issues about automating ethical judgment.

https://doi.org/10.1186/s41073-017-0038-7

Ethics governance

Current research on this topic:
Not much!

Development of Application-Specific Large Language Models to Facilitate Research Ethics Review

Sebastian Porsdam Mann,[1,2,3] Joel Jiehao Seah,[3]
Stephen R. Latham,[4] Julian Savulescu,[3,6*] Mateo Aboy,[5] Brian D. Earp[3,6]

1. Centre for Advanced Studies in Bioscience Innovation Law (CeBIL), Faculty of Law, University of Copenhagen, Karen Blixens Pl. 16, 2300 Copenhagen, Denmark
2. Faculty of Law, University of Oxford, St Cross Building St. Cross Rd, Oxford OX1 3UL, United Kingdom.
3. Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Dr, #02-03 MD 11, Singapore 117597.
4. Yale Interdisciplinary Center for Bioethics, Yale University, 238 Prospect St, New Haven, CT 06511, USA.
5. Centre for Law, Medicine, and Life Sciences (LML) & Centre for Intellectual Property and Information Law (CIPIL), Faculty of Law, University of Cambridge, 5 West Rd, Cambridge CB3 9DP, United Kingdom
6. Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, 16-17 Saint Ebbe's St, Oxford OX1 1PT, United Kingdom.

- Authors (Mann et al.) observe inconsistencies, delays, and inefficiencies in IRB workflows.
- Authors recommend ***IRB-specific LLMs***.
- Authors mention concerns of overreliance and opacity.

Client information

Client information

EthicAlly

EthicAlly is a research project to explore **how AI can support research ethics planning and review** in research with human participants.

We have developed a **prototype powered by Anthropic's Claude.ai*** to demonstrate the potential of this approach.

Our aim is to create a **not-for-profit resource to provide ethics support and assessment** to researchers around the world, including in underressourced regions without research ethics infratsructure.

EthicAlly is currently in **beta testing**: we would love to hear your thougths, critique and suggestions!

Test the system at www.ethically.info and **let us know what you think!**

Want to get involved? We are **looking for collaborators!** Get in touch at contact@ethically.info

*Disclosure: we are not affiliated with Anthropic and do not receive any fianncial or non-financial benefit from them or another commercial sponsor.

https://ethically.info

EthicAlly BETA

Welcome!

EthicAlly is a research ethics support system for the social sciences and humanities, powered by AI. Enter your research proposal and receive a free structured report outlining potential ethical issues and making suggestions to address them. The report is generated by Claude Sonnet 4, a Large Language Model made by Anthropic. Please note that EthicAlly cannot review clinical research.

CAUTION: EthicAlly is a research app and still under development. It can make mistakes. Always seek ethical advice from your institutional research ethics committee or equivalent.

Beta Testing

EthicAlly is currently in beta testing. We're actively improving the system and would love your feedback! Please contact us at contact@ethically.info with your experience, suggestions, or any issues you encounter.

Field of Research *

Computer Science

TU/e

Country/Region *

Netherlands

Research Proposal *

Mitigating Generative AI-Driven Cybersecurity Threats: A Framework for Enhancing Information Security Compliance

1. Introduction

The rapid evolution of Generative Artificial Intelligence (GenAI) has introduced a paradigm shift in the cybersecurity landscape. Once reliant on template-based security policies and compliance frameworks, organizations now face a new class of threats that exploit the generative capabilities of AI to craft sophisticated

Data Processing Notice: Your research proposal will be sent to Anthropic's Claude AI for analysis. Neither the EthicAlly app nor Anthropic retain your data after generating the review. EthicAlly's data processing is automated - no humans review your proposal. For details on Anthropic's data handling, see their privacy policy.

☐ I understand that my research proposal will be processed by AI systems located outside the EU (Anthropic's Claude API). I confirm that I have removed any personally identifying information about research participants and other third parties from my proposal. *

Generate Ethics Review

Note: If you do not receive a response, Claude's servers may be busy - please try again in a few minutes.

✅ Ethics review completed successfully!

Ethics Review Report

DISCLAIMER: This ethics review is generated by an artificial intelligence system for research and educational purposes only. While this analysis applies established ethical frameworks and guidelines, it cannot replace human ethical oversight, institutional review board (IRB) approval, or professional ethics consultation.

1. Summary Assessment

This computer science research proposal presents a legitimate academic investigation into cybersecurity frameworks with appropriate methodological rigor. The study involves organizational participation and expert interviews but lacks critical detail regarding human participant protections, informed consent procedures, and data management protocols required for ethical compliance.

2. Compliance Analysis

Nuremberg Code: While primarily applicable to medical research, the principles of voluntary consent and participant welfare are relevant. The proposal mentions "informed consent" but provides insufficient detail about consent procedures.

Belmont Report:
- *Respect for Persons:* Inadequately addressed - no detailed consent process, unclear participant autonomy protections

References

Nickel, P.J. “Trust in medical artificial intelligence: a discretionary account.” *Ethics and Information Technology* 24 (2022). DOI: 10.1007/s10676-022-09630-5

Nickel, P.J. “The prospect of artificial intelligence supported ethics review.” *Ethics & Human Research* 46, 6 (2024): 25-28. DOI: 10.1002/eahr.500230

Porsdam Mann, S., et al. “Development of Application-Specific Large Language Models to Facilitate Research Ethics Review.” *arXiv e-prints* (2025): arXiv-2501.

Topic 3: AI tool to support people with eating disorder symptoms

- Clinically tested interactive tool
- Currently the chat interface is not AI-based. The question is how an AI tool could be developed to support patients, family members, and/or the public on this topic

First steps of the assignment

- Gradually form a group, preferably a group with multidisciplinary perspectives and multiple academic and cultural backgrounds. You can’t finalize this today, so stay open about it.
- Learn about the ***target context.*** What are the problems and challenges that practitioners and other stakeholders have in that context?
- Do background research on academic and grey literature regarding these problems.
- Avoid “techno-solutionism”: “the problem definition follows the solution and is twisted to fit the solution, rather than the other way around” (Siffels & Sharon 2024, p. 3)

Orientation activity

- See the canvas survey, linkable from Modules

Please introduce yourself to at least four other people during the self-study period on 2 September, and answer the survey questions.

Please complete the survey.