

The social dilemma in AI development

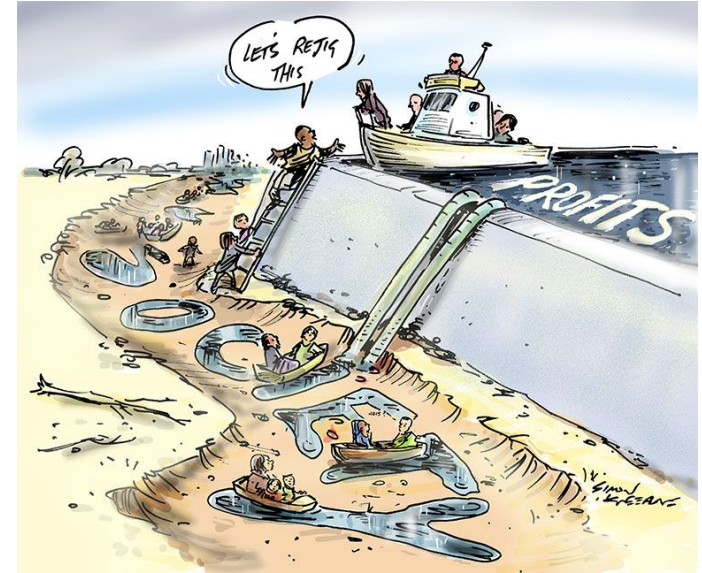
inga.strumke@ntnu.no

What is a social dilemma?

A social dilemma ...

is a situation that tempts a person to seek a selfish gain by neglecting the interests of others





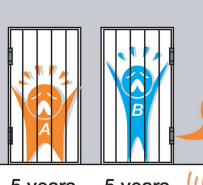


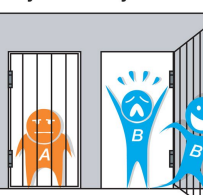
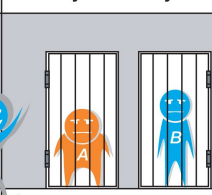
Also referred to as a 'collective action problem', it is a decision-making problem faced when the **interests of the collective conflict with the interests of the individual(s)** making a decision.



Examples?

Collective action everywhere :/

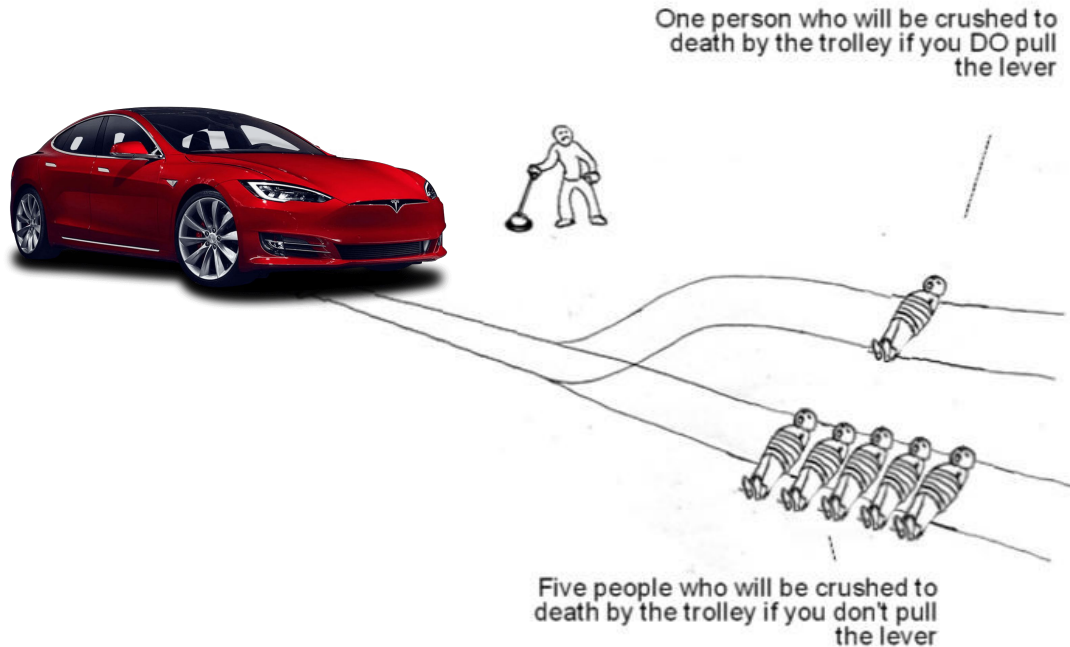
Prisoners' dilemma

		prisoner B	
		confess 	remain silent 
prisoner A 	confess 	 5 years 5 years	 0 year 20 years
	remain silent 	 20 years 0 year	 1 year 1 year



The social dilemma in AI development?

The social dilemma in AI development



This isn't the social dilemma you're looking for

The social dilemma in AI development

The demand for AI systems increases

The demand for *ethical development* increases

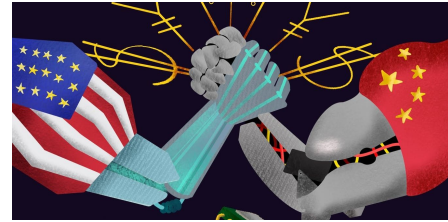
The number of ethical guidelines and white papers increases

Still, the number of unethical use cases of AI accelerates

Why?

\$15.7tr

Potential contribution to the global economy by 2030 from AI



The social dilemma in AI development

Actual example:

A company develops an AI tool to be used to guide hiring decisions.

The social dilemma in AI development

Actual example:

A company develops an AI tool to be used to guide hiring decisions.

After the product has reached a certain stage, a developer identifies ethical challenges, including recognising that the tool is discriminatory against minorities.

The social dilemma in AI development

Actual example:

A company develops an AI tool to be used to guide hiring decisions.

After the product has reached a certain stage, a developer identifies ethical challenges, including recognising that the tool is discriminatory against minorities.

Avoiding this discrimination may require decreasing the performance of the product.

The social dilemma in AI development

Actual example:

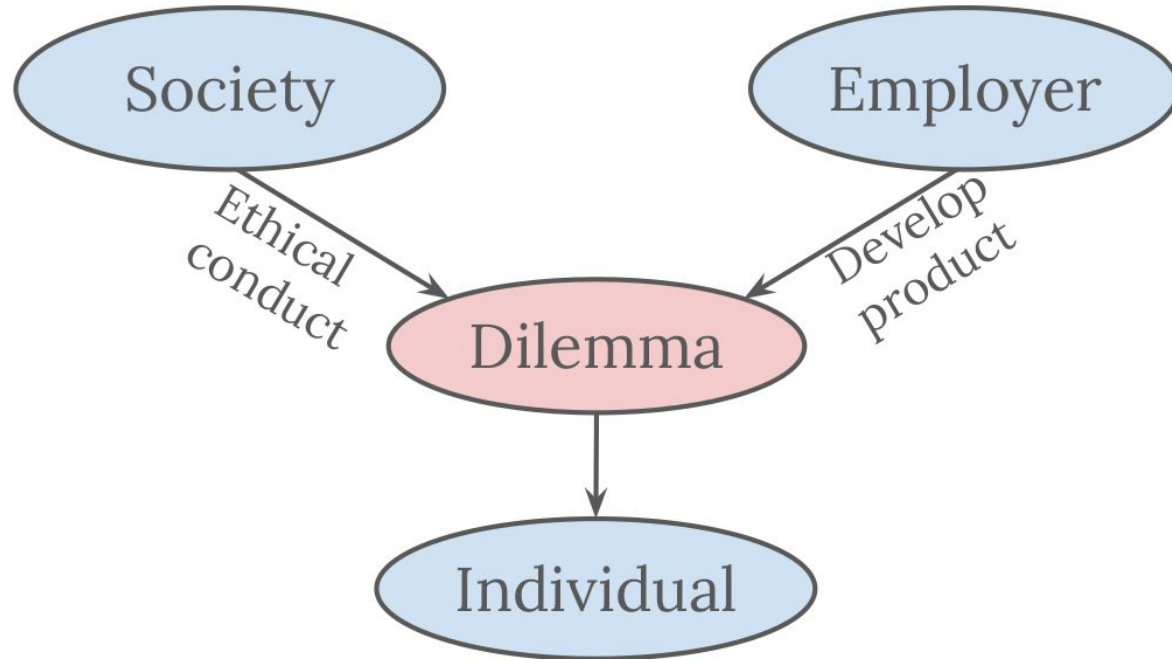
A company develops an AI tool to be used to guide hiring decisions.

After the product has reached a certain stage, a developer identifies ethical challenges, including recognising that the tool is discriminatory against minorities.

Avoiding this discrimination may require decreasing the performance of the product.

What should the developer do?

The social dilemma in AI development



Solution?

Societal awareness?

Individual responsibility?



Google fires second AI ethics leader as dispute over research, diversity grows

How one employee's exit shook Google and the AI industry

Individual moral.

Societal awareness?

Individual responsibility?



Google fires second AI ethics leader as dispute over research, diversity grows

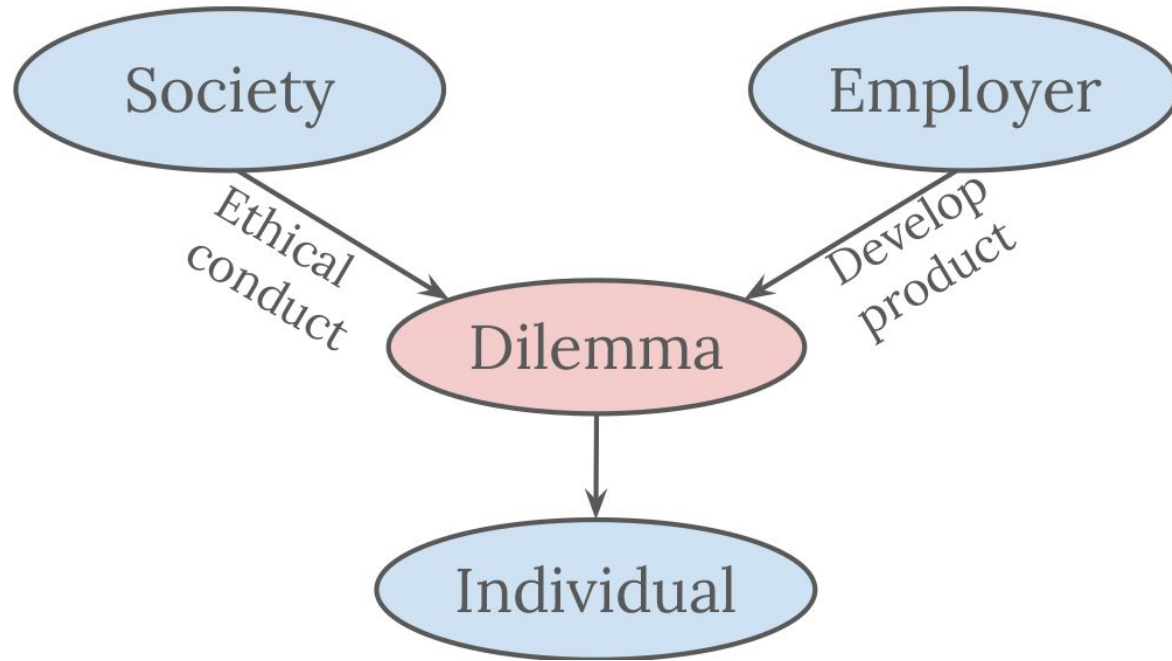
How one employee's exit shook Google and the AI industry

Meanwhile, elsewhere...

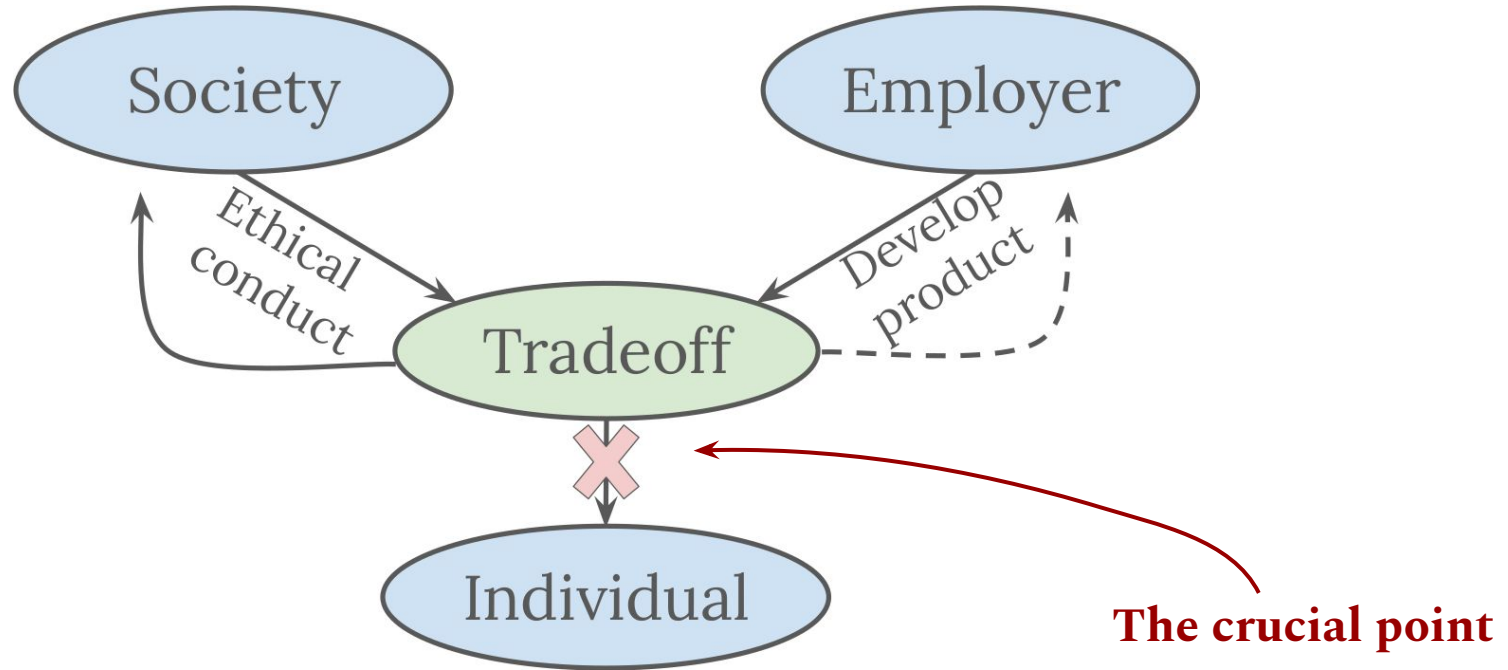


We have little reason to expect that individual moral will solve the social dilemma in AI development

Which **connection** makes this a social dilemma?



How to solve the social dilemma in AI development?





AI and Ethics

The social dilemma in artificial intelligence development and why we have to solve it

Inga Strömke, Marija Slavkovic, Vince I. Madai

While the demand for ethical artificial intelligence (AI) systems increases, the number of unethical uses of AI accelerates, even though there is no shortage of ethical guidelines. We argue that a possible underlying cause for this is that AI developers face a social dilemma in AI development ethics, preventing the widespread adaptation of ethical best practices. We define the social dilemma for AI development and describe why the current crisis in AI development ethics cannot be solved without relieving AI developers of their social dilemma. We argue that AI development must be professionalised to overcome the social dilemma, and discuss how medicine can be used as a template in this process.

Society has solved the social dilemma for medical personnel.

We suggest professionalising AI development and use medicine as a template for professional ethics, based on similar potential impact on society.

I'm a medical doctor;
not an ethicist

I'm an AI developer;
not an ethicist

Medicine impacts people's
lives and society at large.

AI impacts people's lives
and society at large.

I have a professional ethos
for guidance and protection

I have **no** professional
ethos or protection



Parallel to medicine

Large impact on society

Challenging ethical dilemmas

Domain experts; not ethicists

Medical professionals have ethical codes of conduct; AI professionals do not

Should AI development become professionalised?



Professions

A **profession** is a disciplined group of individuals who adhere to **ethical standards** and who hold themselves out as, and are accepted by the public as possessing **special knowledge** and skills in a widely recognised body of learning derived from **research, education** and **training** at a high level, and who are prepared to **apply this knowledge and exercise these skills in the interest of others**.

It is inherent in the definition of a profession that a **code of ethics governs the activities of each profession**. Such codes require behaviour and practice **beyond the personal moral obligations of an individual**.

Australian Council of Professions, 2003

Professionals

A professional is a member of a profession. Professionals are **governed by codes of ethics** and profess commitment to competence, integrity and morality, altruism and the **promotion of the public good within their expert domain**. Professionals are accountable to those they serve and to **society**.

- Evetts, J., 'Sociological Analysis of Professionalism: Past, Present and Future', Comparative Sociology 10, 2011
- Freidson, E., 'Professionalism: The Third Logic', Polity Press, London, 2001

New decisions

PNAS

PNAS

PNAS

PNAS



Facebook language predicts depression in medical records

Johannes C. Eichstaedt^{a,1,2}, Robert J. Smith^{b,1}, Raina M. Merchant^{b,c}, Lyle H. Ungar^{a,b}, Patrick Crutchley^{a,b}, Daniel Preoțiu-Pietro^a, David A. Asch^{b,d}, and H. Andrew Schwartz^e

^aPositive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104; ^bPenn Medicine Center for Digital Health, University of Pennsylvania, Philadelphia, PA 19104; ^cDepartment of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, PA 19104; ^dThe Center for Health Equity Research and Promotion, Philadelphia Veterans Affairs Medical Center, Philadelphia, PA 19104; and ^eComputer Science Department, Stony Brook University, Stony Brook, NY 11794

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved September 11, 2018 (received for review February 26, 2018)

Depression, the most prevalent mental illness, is underdiagnosed and undertreated, highlighting the need to extend the scope of current screening methods. Here, we use language from Facebook posts of consenting individuals to predict depression recorded in electronic medical records. We accessed the history of Facebook statuses posted by 683 patients visiting a large urban academic emergency department, 114 of whom had a diagnosis of depression in their medical records. Using only the language preceding their first documentation of a diagnosis of depression, we could identify depressed patients with fair accuracy [area under the curve (AUC) = 0.69], approximately matching the accuracy of screening surveys benchmarked against medical records. Restricting Facebook data to only the 6 months immediately preceding the first documented diagnosis of depression yielded a higher prediction accuracy (AUC = 0.72) for those users who had sufficient Facebook data.

Significant prediction of future depression status was possible as far as 3 months before its first documentation. We found that language predictors of depression include emotional (sadness), interpersonal (loneliness, hostility), and cognitive (preoccupation with the self, rumination) processes. Unobtrusive depression assessment through social media of consenting individuals may become feasible as a scalable complement to existing screening and monitoring procedures.

the diagnosis of depression, which prior research has shown is feasible with moderate accuracy (15). Of the patients enrolled in the study, 114 had a diagnosis of depression in their medical records. For these patients, we determined the date at which the first documentation of a diagnosis of depression was recorded in the EMR of the hospital system. We analyzed the Facebook data generated by each user before this date. We sought to simulate a realistic screening scenario, and so, for each of these 114 patients, we identified 5 random control patients without a diagnosis of depression in the EMR, examining only the Facebook data they created before the corresponding depressed patient's first date of a recorded diagnosis of depression. This allowed us to compare depressed and control patients' data across the same time span and to model the prevalence of depression in the larger population (~16.7%).

Results

Prediction of Depression. To predict the future diagnosis of depression in the medical record, we built a prediction model by using the textual content of the Facebook posts, post length, frequency of posting, temporal posting patterns, and demographics (*Materials and Methods*). We then evaluated the performance of this model by comparing the probability of depression estimated by our algorithm

PSYCHOLOGICAL AND
COGNITIVE SCIENCES

Old possibilities with new potential

Dual use of artificial-intelligence-powered drug discovery

[Fabio Urbina](#), [Filippa Lentzos](#), [Cédric Invernizzi](#) & [Sean Ekins](#) 

[Nature Machine Intelligence](#) **4**, 189–191 (2022) | [Cite this article](#)

100k Accesses | 2656 Altmetric | [Metrics](#)

An international security conference explored how artificial intelligence (AI) technologies for drug discovery could be misused for de novo design of biochemical weapons. A thought experiment evolved into a computational proof.

“

... our model generated 40,000 molecules that scored within our desired threshold. In the process, the AI designed (...) many known chemical warfare agents that we identified through visual confirmation with structures in public chemistry databases. Many new molecules were also designed (...) These new molecules were predicted to be more toxic, based on the predicted LD50 values, than publicly known chemical warfare agents. This was unexpected because the datasets we used for training the AI did not include these nerve agents.

”

Ethics (and possibly legislation) might end up being all we have

Planting Undetectable Backdoors in Machine Learning Models

Shafi Goldwasser
UC Berkeley

Michael P. Kim
UC Berkeley

Vinod Vaikuntanathan
MIT

Or Zamir
IAS

Abstract

Given the computational cost and technical expertise required to train machine learning models, users may delegate the task of learning to a service provider. Delegation of learning has clear benefits, and at the same time raises *serious concerns of trust*. This work studies possible abuses of power by untrusted learners.

We show how a malicious learner can plant an *undetectable backdoor* into a classifier. On the surface, such a backdoored classifier behaves normally, but in reality, the learner maintains a mechanism for changing the classification of any input, with only a slight perturbation. Importantly, without the appropriate “backdoor key,” the mechanism is hidden and cannot be detected by any computationally-bounded observer. We demonstrate two frameworks for planting undetectable backdoors, with incomparable guarantees.

Thank you:)

inga.strumke@ntnu.no