

DIGNITY EXPERIMENT



DIGNITY EXPERIMENT

Odds of a single message turning into a conversation

		number DISPLAYED to them		
		30% match	60% match	90% match
ACTUAL compatibility of users	30% match	10%	16%	17%
	60% match	13%	13%	16%
	90% match	16%	17%	20%

As you can see, the ideal situation is the lower right: to both be *told* you're a good match, and at the same time actually *be* one. OkCupid definitely works, but that's not the whole story. And if you have to choose only one or the other, the mere myth of compatibility works just as well as the truth.

DIGNITY EXPERIMENT

Reactions? (If you're an OKCupid user...)

Samples from blog's comments:

apb_101 says:

Good to know at any one time here forward I can trust or distrust your Match %s, not knowing if your lab experiment has resurfaced.

an0n says:

@SteveRestless

What if the manipulation is used for what you believe does good for the person. That is what I am forced to believe is the reason, people end up manipulating people for money or good in their own heart.

DIGNITY EXPERIMENT

What *is* human dignity?

What is the value of dignity?

How can it be described, or measured? (What are the “units” of dignity?)

OUTLINE

Dignity experience: Experiments on users

Dignity definition

Dilemma: Dignity + psychological health (+ Mimetic AI)

AI Weapons dilemma

Dignity versus happiness/utilitarianism

Who gets to be dignified?

DEFINITION DIGNITY

Definitions

- Conceptual
- Philosophical negative/protective
- Philosophical positive/enabling

The concept of human dignity has been depicted through an evocative scene, portraying a diverse group of people unified in strength and hope as they stand together at sunrise. This image symbolizes the inherent worth and potential within each individual and the collective power of humanity, highlighting themes of respect, equality, and resilience.

ChatGPT 2024



DEFINITION DIGNITY

Conceptual

All humans have inherent / inalienable / indivisible value

Have it, cannot lose it, cannot compare it against others...

DEFINITION DIGNITY (IMMANUEL KANT)

Philosophical negative/protective

The right to be treated as an end-in-yourself, and not only as a means to someone else's ends.

- Kant, 1785: *Groundwork Metaphysics of Morals*

The right to not be used as only a tool or instrument

The right to not be exploited/manipulated

Case: *If I hire you to write me some code, and I buy a computer for you to work with, is there a difference between how I treat you and how I treat the computer?*

DEFINITION DIGNITY

Philosophical negative/protective

The right to be not disposable/replaceable
= not commodified



DEFINITION DIGNITY

Philosophical **positive/enabling**

The right to take responsibility for your own actions,
not suffer condescension, not treated as a baby
= right to accountability

"A baby steals toy from another baby,
and the victim baby cries: OpenAI"

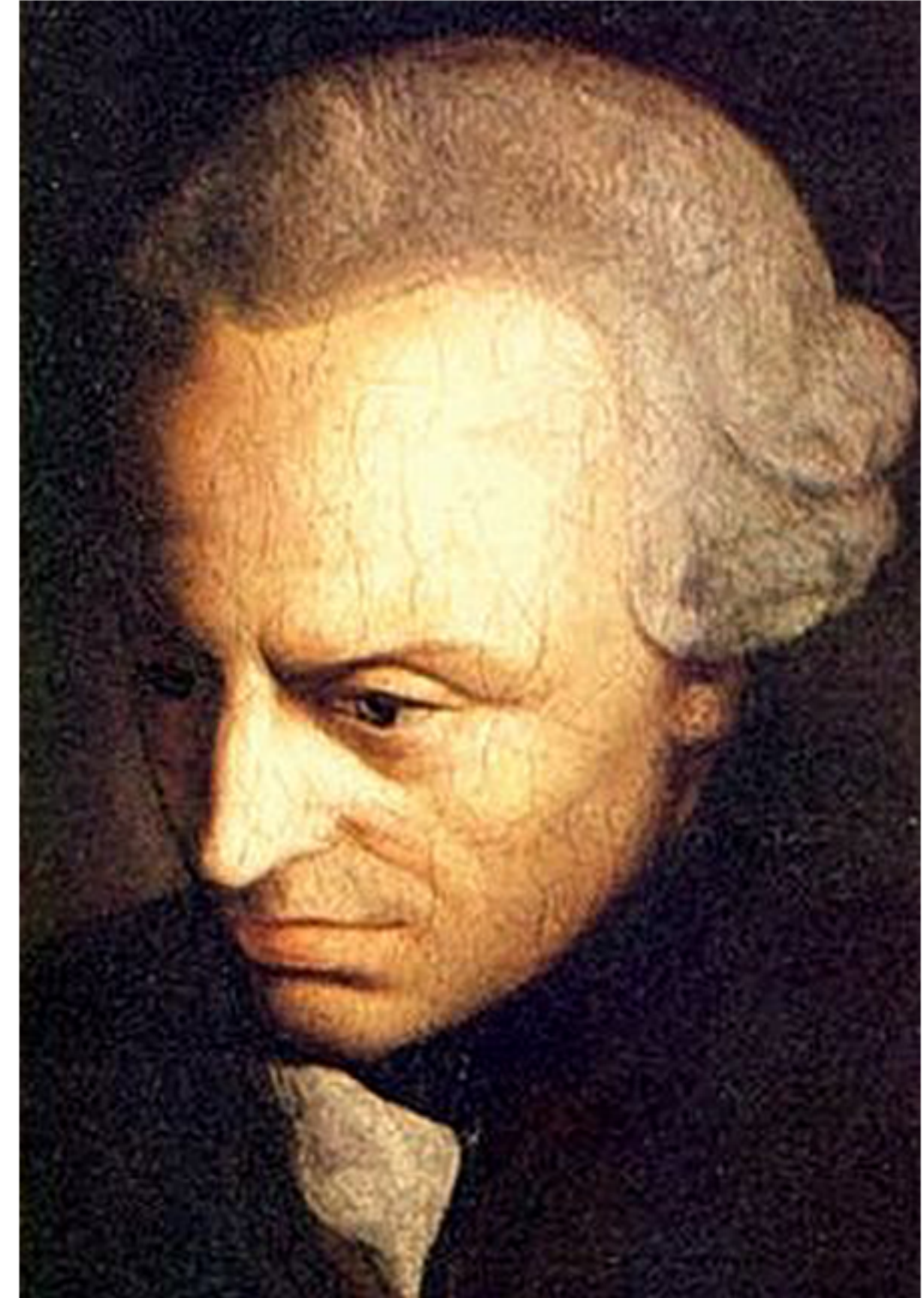


DEFINITION DIGNITY

Philosophical positive/enabling

“Even if a civil society were to dissolve itself by common agreement of all its members (for example, if the people inhabiting an island decided to separate and disperse themselves around the world), the last murderer remaining in prison must first be executed, so that everyone will duly receive what his actions are worth.”

– Kant, *Metaphysics of Morals*, Part II, *The Science of Right*



Question: *Who benefits from the execution ethically, who wants the execution?*

DEFINITION DIGNITY

All humans have inherent / inalienable / indivisible value

We are not commodities, not replaceable

We are accountable

Case: Is the user *dignified* by this interaction?

Terms and Conditions

Terms And Conditions * ☐ I have read and agree to the [Terms and Conditions](#).

Register

OUTLINE

Dignity experience: Experiments on users

Dignity definition

- Treated as end, not only as means (not commodified)

- The right of accountability

Dilemma: Dignity + psychological health (+ Mimetic AI)

AI Weapons dilemma

Dignity versus happiness/utilitarianism

Who gets to be dignified?

DIGNITY DILEMMA ELEMENTS: WOEBOT+MIMETIC AI = IMPERSONATION

“Woebot feels more like a friend than an app.”



Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess

Reid McIlroy-Young
University of Toronto

Russell Wang
University of Toronto

Siddhartha Sen
Microsoft Research

Jon Kleinberg
Cornell University

Ashton Anderson
University of Toronto

Abstract

The advent of machine learning models that surpass human decision-making ability in complex domains has initiated a movement towards building AI systems that interact with humans. Many building blocks are essential for this activity, with a central one being the algorithmic characterization of human behavior. While much of the existing work focuses on aggregate human behavior, an important long-range goal is to develop behavioral models that specialize to individual people and can differentiate among them. To formalize this process, we study the problem of *behavioral stylometry*, in which the task is to identify a decision-maker from their decisions alone. We present a transformer-based approach to behavioral stylometry in the context of chess, where one attempts to identify the player who played a set of games. Our method operates in a few-shot classification framework, and can correctly identify a player from among thousands of candidate players with 98% accuracy given only 100 labeled games. Even when trained on amateur play, our method generalises to out-of-distribution samples of Grandmaster players, despite the dramatic differences

SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION

Akari Asai[†], Zeqiu Wu[†], Yizhong Wang^{†§}, Avirup Sil[‡], Hannaneh Hajishirzi^{†§}
[†]University of Washington [§]Allen Institute for AI [‡]IBM Research AI
{akari, zeqiuwu, yizhongw, hannaneh}@cs.washington.edu, avi@us.ibm.com

ABSTRACT

Despite their remarkable capabilities, large language models (LLMs) often produce responses containing factual inaccuracies due to their sole reliance on the parametric knowledge they encapsulate. Retrieval-Augmented Generation (RAG), an ad hoc approach that augments LMs with retrieval of relevant knowledge, decreases such issues. However, indiscriminately retrieving and incorporating a fixed number of retrieved passages, regardless of whether retrieval is necessary, or passages are relevant, diminishes LM versatility or can lead to unhelpful response generation. We introduce a new framework called **Self-Reflective Retrieval-Augmented Generation (SELF-RAG)** that enhances an LM’s quality and factuality through retrieval and self-reflection. Our framework trains a single arbitrary LM that adaptively retrieves passages on-demand, and generates and reflects on retrieved passages and its own generations using special tokens, called *reflection* tokens. Generating reflection tokens makes the LM controllable during the inference phase, enabling it to tailor its behavior to diverse task requirements. Experiments show that SELF-RAG (7B and 13B parameters) significantly outperforms state-of-the-art LLMs and retrieval-augmented models on a diverse set of tasks. Specifically, SELF-RAG outperforms ChatGPT and retrieval-augmented Llama2-chat on Open-domain QA, reasoning and fact verification tasks, and it shows significant gains in improving factuality and citation accuracy for long-form generations relative to these models.¹

1 INTRODUCTION

State-of-the-art LLMs continue to struggle with factual errors (Mallen et al., 2023; Min et al., 2023) despite their increased model and data scale (Ouyang et al., 2022). Retrieval-Augmented Generation (RAG) methods (Figure 1 left; Lewis et al. 2020; Guu et al. 2020) augment the input of LLMs with relevant retrieved passages, reducing factual errors in knowledge-intensive tasks (Ram et al.

DIGNITY DILEMMA #1: WOEBOT + MIMETIC AI



If it works better when she believes her own doctor/nurse/friend is on the other end of the phone, do you tell her it is a machine?



What if *you* were elderly?

Would you prefer your own children
to lie to you if it would help alleviate
your depression?

Counterfeit language argument against:

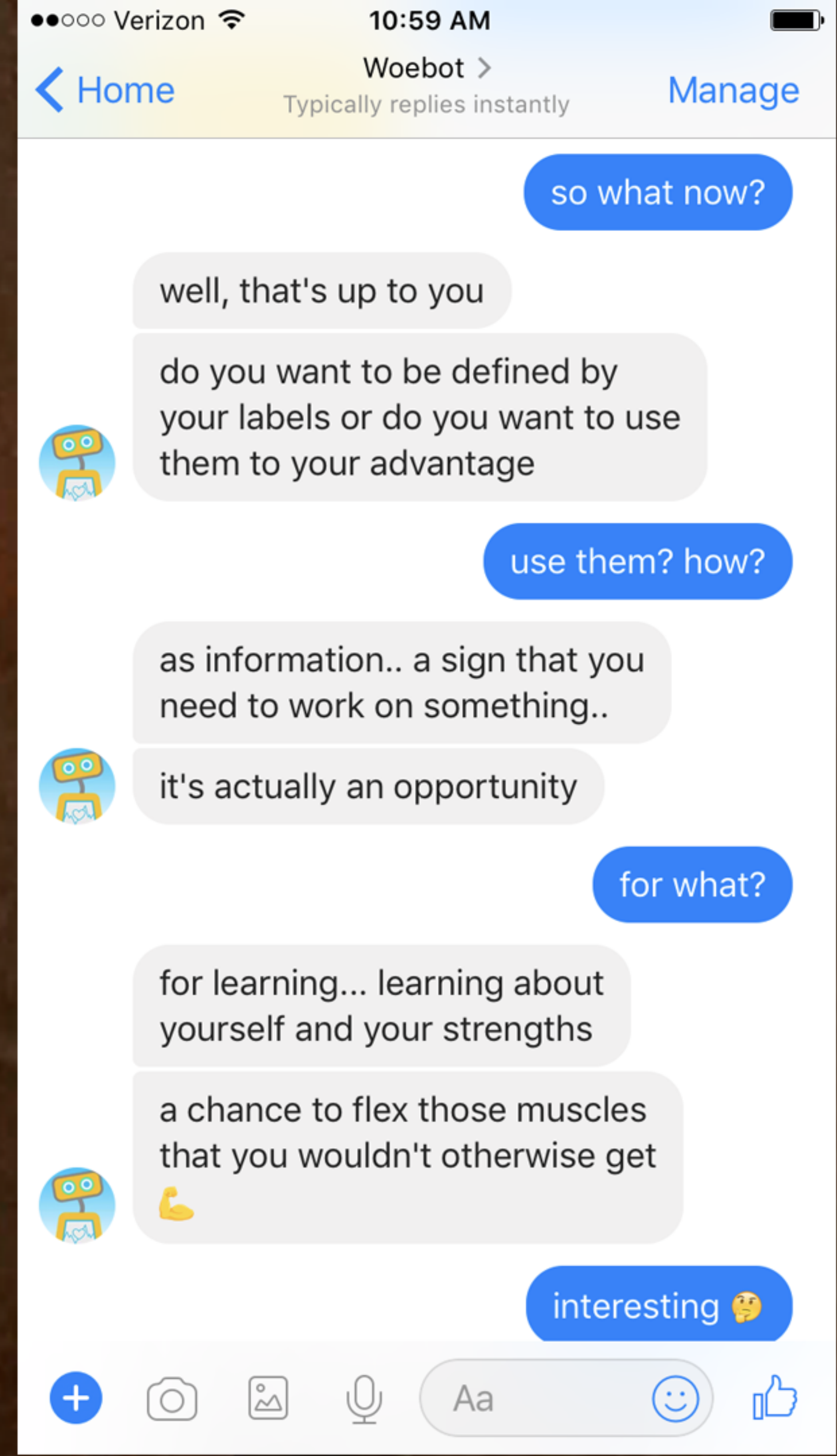
Does not upvalue machines, but degrades people.
Counterfeiting = genuine currency devalues.

– Frank Pasquale and others

True?

How could a developer counter technically?

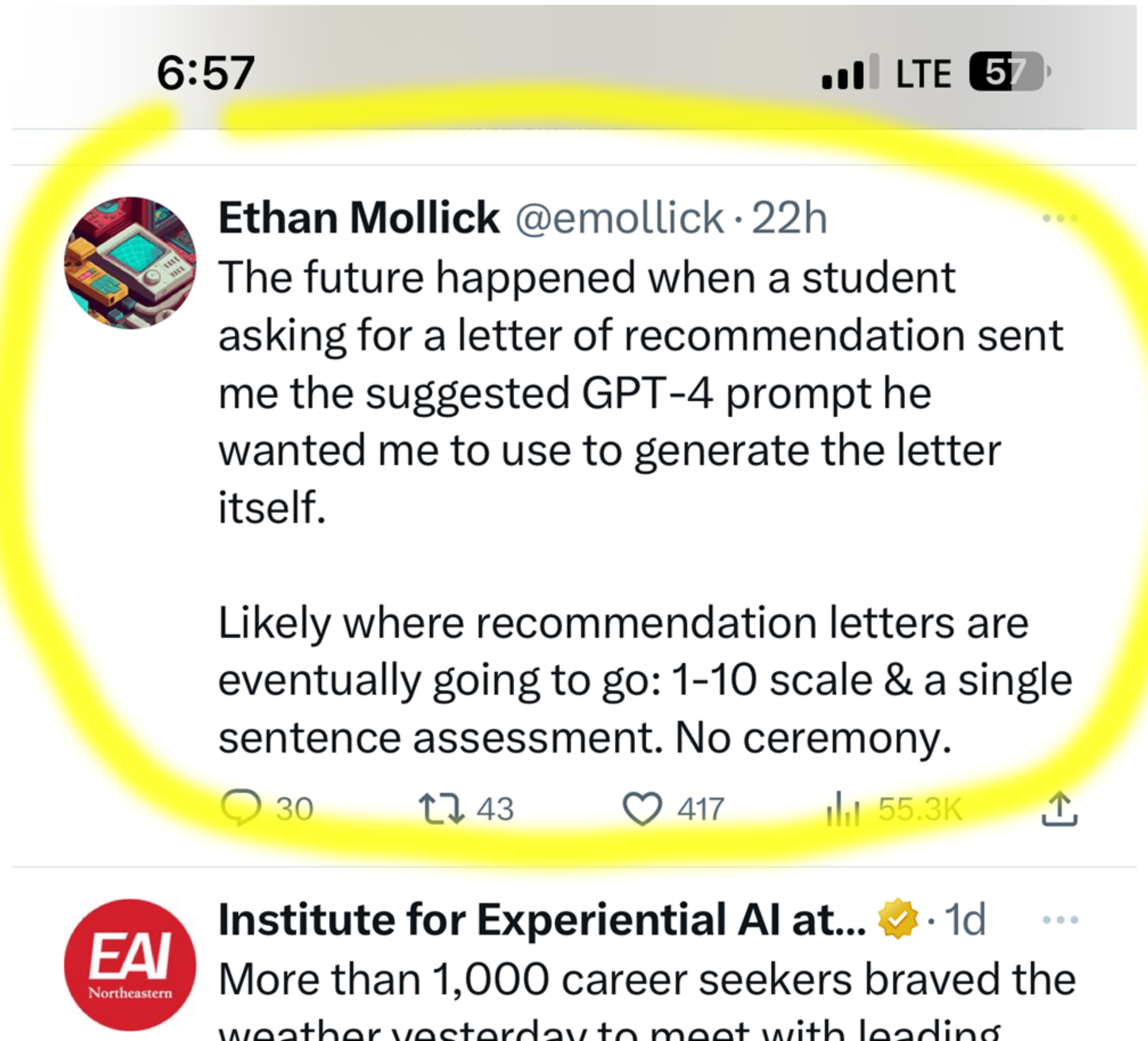
How could a developer counter ethically?



DIGNITY DILEMMA #1: ANOTHER EXAMPLE...

Counterfeit:

Do machines come up to our level,
or lower us down to theirs?



DIGNITY DILEMMA #1: WOEBOT + MIMETIC AI



Balanced values argument for:

Dignity in conflict with:
Autonomy?
Biological health
(Is dignity worth death)?
Happiness?

DIGNITY DILEMMA #2, PRACTICAL ETHICS

THE 24/7 DOCTOR

Personal doctor/therapist on the phone on demand, 24/7/365. Doctor is avatar, but the personalization is indistinguishable from human.

Assuming a medical benefit to the deception (treating minor depression for example, or gestational diabetes), here are 2 beneficial strategies:

- 1) Tell the patient the general truth, but randomize the human and AI doctor so the patient is never certain which one is on the phone. Patient must be prohibited from asking. (An aggressive placebo strategy...?)

- 2) Go all in and hide the AI stand-in from the patient

(Note: What if the patient constantly suspects the real doctor is AI and net result is worse care...?)

Which would you do? Is there another solution?

DIGNITY DILEMMA #3, BE RIGHT BACK (BLACK MIRROR)

POST MORTEM DIGITAL TWIN

"Ex" after death

OTHER SCENARIOS...

"Ex" after relocated temporarily to a distant time zone

"Ex" after leaves and finds someone else

"Ex" when used by someone who never knew him/her

Does passing time change the dignity consideration?

Days? Years? Centuries?

OUTLINE

Dignity experience: Experiments on users

Dignity definition

- Treated as end, not only as means

- The right of accountability

Dilemma: Dignity + psychological health (+ Mimetic AI)

- Manipulation & justifying manipulation

AI Weapons dilemma

Dignity versus happiness/utilitarianism

Who gets to be dignified?

ARE AI WEAPONS DIGNIFIED?

Does dignity ethics preclude or sanction AI in war?

2 debates



ARE AI WEAPONS DIGNIFIED? DIGNITY OF HUMAN LIFE DEBATE

No:

Decisions involving human lives should remain human judgment as only humans understand the dignity of life.

Makes killing counterfeit, like a video game - makes life cheaper...

But:

Is there any legitimate difference between AI and rifles? Landmines?



ARE AI WEAPONS DIGNIFIED? ACCOUNTABILITY / RESPONSIBILITY DEBATE

No:

Who – the designer, the operator, the commander, or the AI itself – is responsible for AI war acts? No one knows...

But:

Lives lost without accountability... but also saves lives.



OUTLINE

Dignity experience: Experiments on users

Dignity definition

- Treated as end, not only as means

- The right of accountability

Dilemma: Dignity + psychological health (+ Mimetic AI)

- Manipulation & justifying manipulation

AI Weapons dilemma

Dignity versus happiness/utilitarianism

Who gets to be dignified?

DIGNITY *VERSUS* HAPPINESS/UTILITARIANISM

“Man does not strive for happiness; only the Englishman does that.”

- Nietzsche, *Twilight of the Idols* (1888)

Two *kinds* of comparison between dignity and happiness

- ↗ Dignity versus happiness as countable amounts (what are the metrics...?)
- ↘ Innumerable/indivisible dignity *versus* numerable/divisible happiness
calculation = measurable ethics (Does dignity *force* all or nothing?)

		Odds of a single message turning into a conversation		
		number DISPLAYED to them		
		30% match	60% match	90% match
ACTUAL compatibility of users	30% match	10%	16%	17%
	60% match	13%	13%	16%
	90% match	16%	17%	20%

OUTLINE

Dignity experience: Experiments on users

Dignity definition

- Treated as end, not only as means
- The right of accountability

Dilemma: Dignity + psychological health (+ Mimetic AI)

- Manipulation & Justifying manipulation

AI Weapons dilemma

Dignity versus happiness/utilitarianism

Who gets to be dignified?

WHO GETS TO BE DIGNIFIED?

*Where does dignity come from?

Self-consciousness/Intentions?
Intelligence?
The claim of dignity?

