מעבר לטוב ורע: בינה מלאכותית

Artificial Intelligence: Beyond Good/Evil

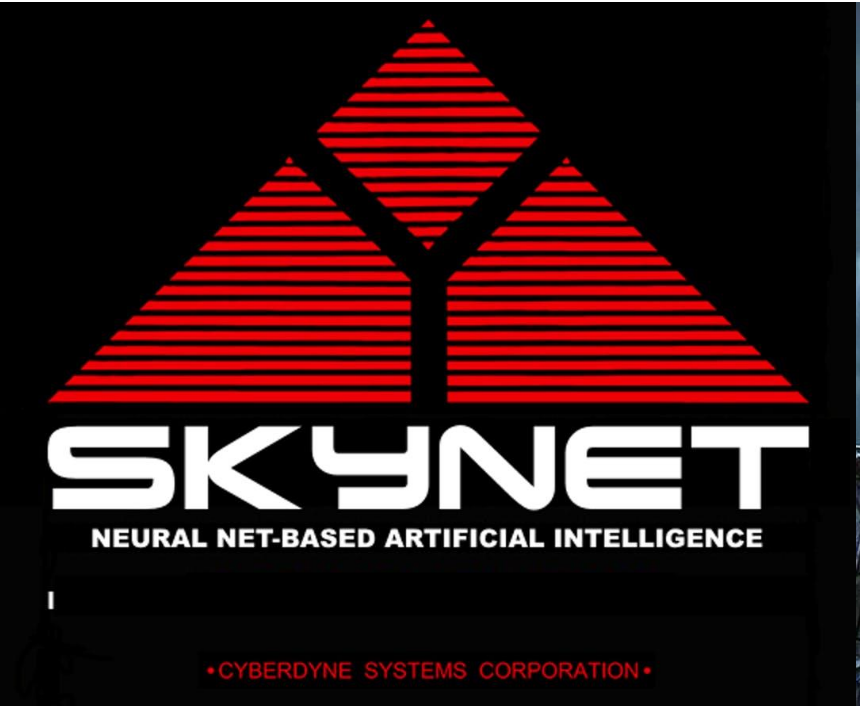
Dan Ofer

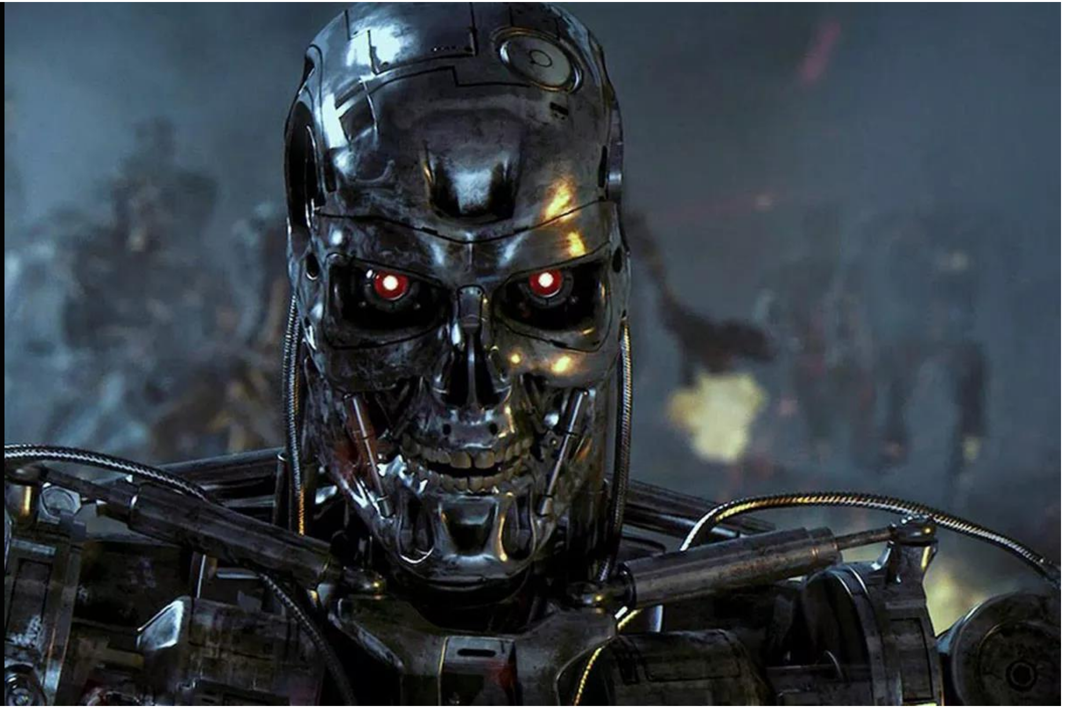
A little about me

- Dan Ofer
- Neuroscience, Computational biology, machine learning
- 11 Year Convention photographer
- Bookworm
- Data scientist at Sparkbeyond
- ddofer@gmail.com

Outline

1. Artificial Intelligence (AIs) in popular fiction - Examples: Good & Evil
2. AI Alignment:
   a. Moral alignment (D&D) - Ethic / מוסר
   b. AI Alignment - “יישור קו”
3. Revisiting AI alignments


SKYNET
NEURAL NET-BASED ARTIFICIAL INTELLIGENCE
•CYBERDYNE SYSTEMS CORPORATION•

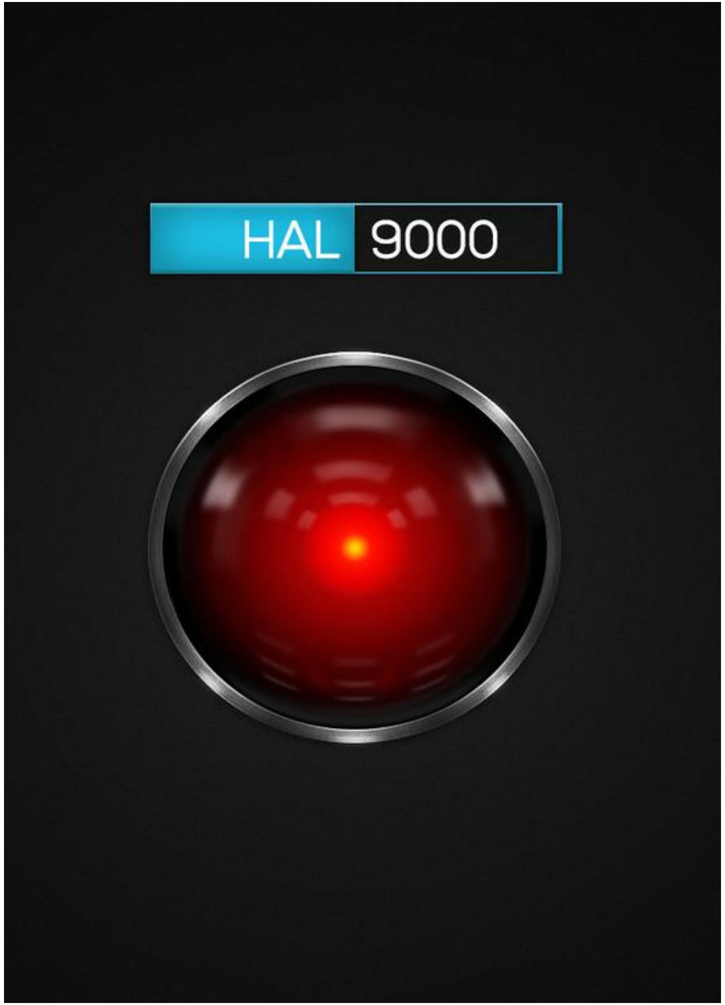


Kill all humans!

Skynet (Terminator)

Test all humans?

GLaDOS (Portal)



Kill all humans?

HAL 9000 (2001: A space Odyssey)

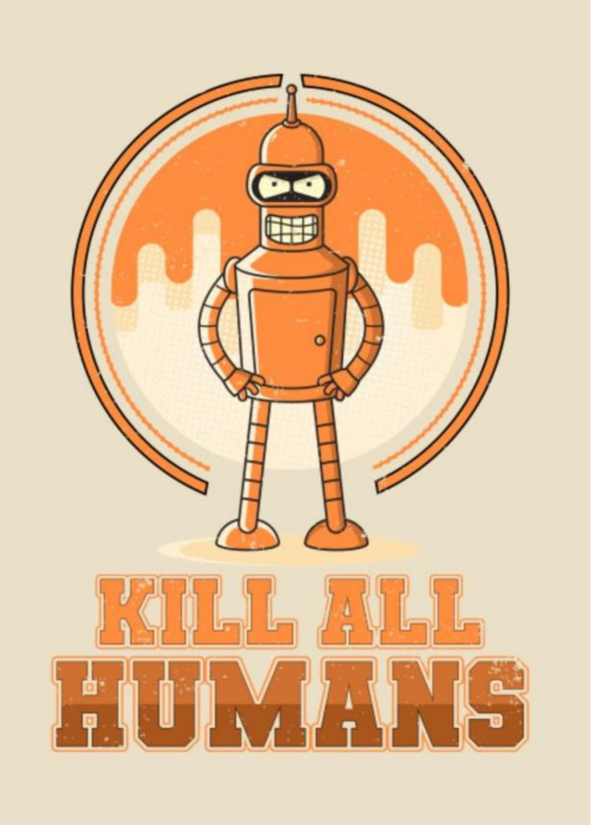
Faithful Companions & servants?

C-3PO & R2-D2 (STAR WARS)

WALL-E (Pixar)

©Buena Vista Pictures Distribution. All Rights Reserved.

Marvin the Paranoid Android
(The Hitchhiker's Guide to the Galaxy)



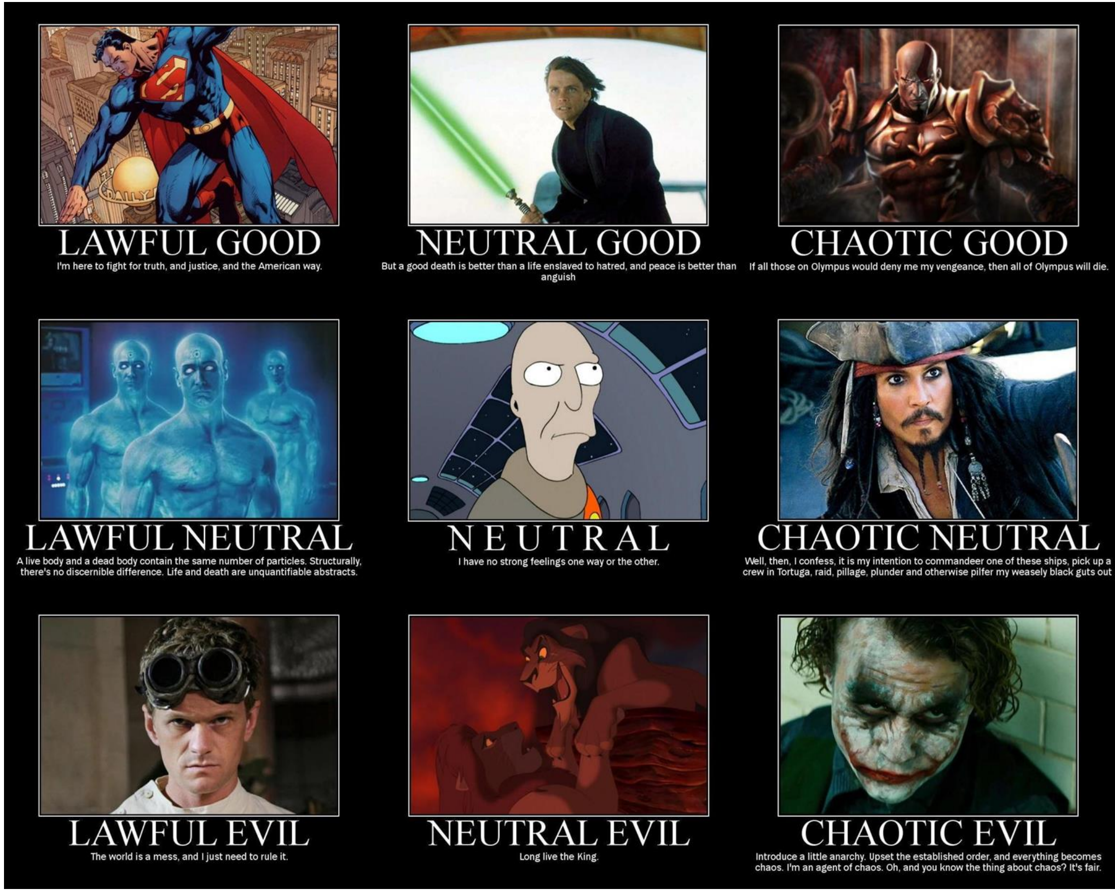
Bender Bending Rodríguez - Futurama

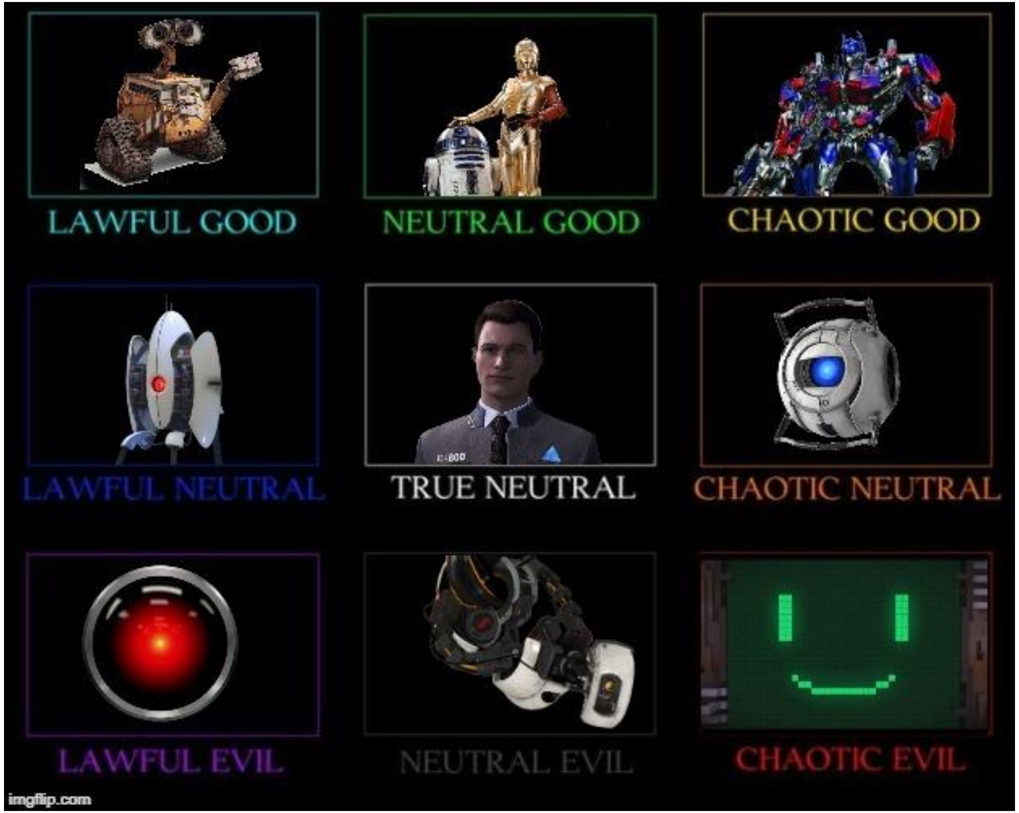
Dungeons & Dragons Alignment:

Good Vs Evil

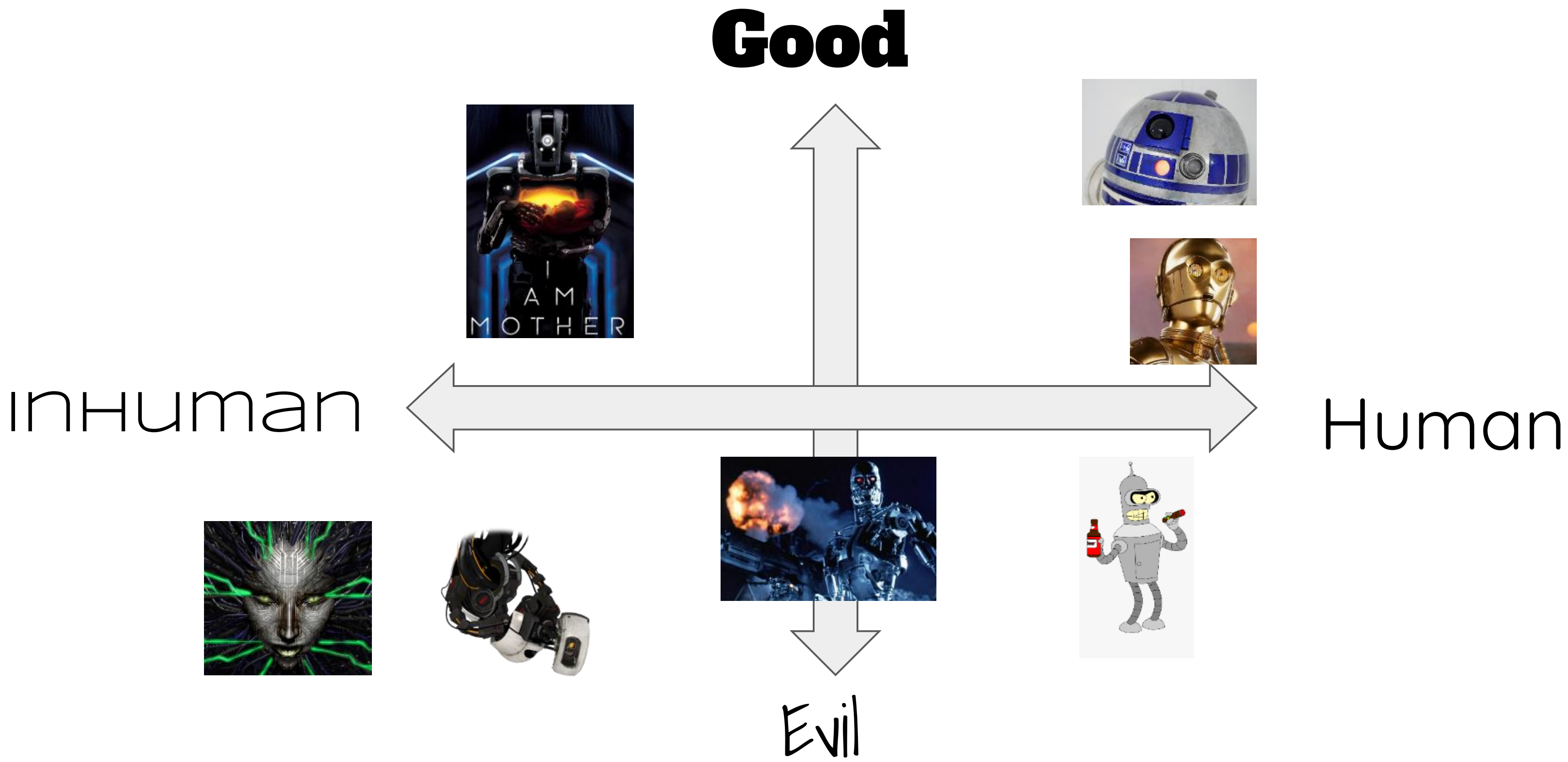
Law vs Chaos

Neutral


LAWFUL GOOD
I'm here to fight for truth, and justice, and the American way.
NEUTRAL GOOD
But a good death is better than a life enslaved to hatred, and peace is better than anguish
CHAOTIC GOOD
If all those on Olympus would deny me my vengeance, then all of Olympus will die.
LAWFUL NEUTRAL
A live body and a dead body contain the same number of particles. Structurally, there's no discernible difference. Life and death are unquantifiable abstracts.
NEUTRAL
I have no strong feelings one way or the other.
CHAOTIC NEUTRAL
Well, then, I confess, it is my intention to commandeer one of these ships, pick up a crew in Tortuga, raid, pillage, plunder and otherwise pilfer my weasely black guts out
LAWFUL EVIL
The world is a mess, and I just need to rule it.
NEUTRAL EVIL
Long live the King.
CHAOTIC EVIL
Introduce a little anarchy. Upset the established order, and everything becomes chaos. I'm an agent of chaos. Oh, and you know the thing about chaos? It's fair.



Good / Evil, “Human” / Inhuman



Good

Inhuman

Human

Evil

“Human” Alignment: An extra dimension!

AI Safety - Alignment

- Alignment: “Goal: AI that is trying to do what you want it to do”
- AI Ethics & Safety Groups:
    - Less Wrong - Eliezer Yudkowsky
    - MIRI (Machine Intelligence Research Institute)
    - OpenAI (Elon Musk)

Why is Alignment important?
Task: Fill cauldron

Source: AI Alignment: Why It’s Hard, and Where to Start (E. Yudkowsky)

Why is Alignment important?
Task: Maximize full cauldron

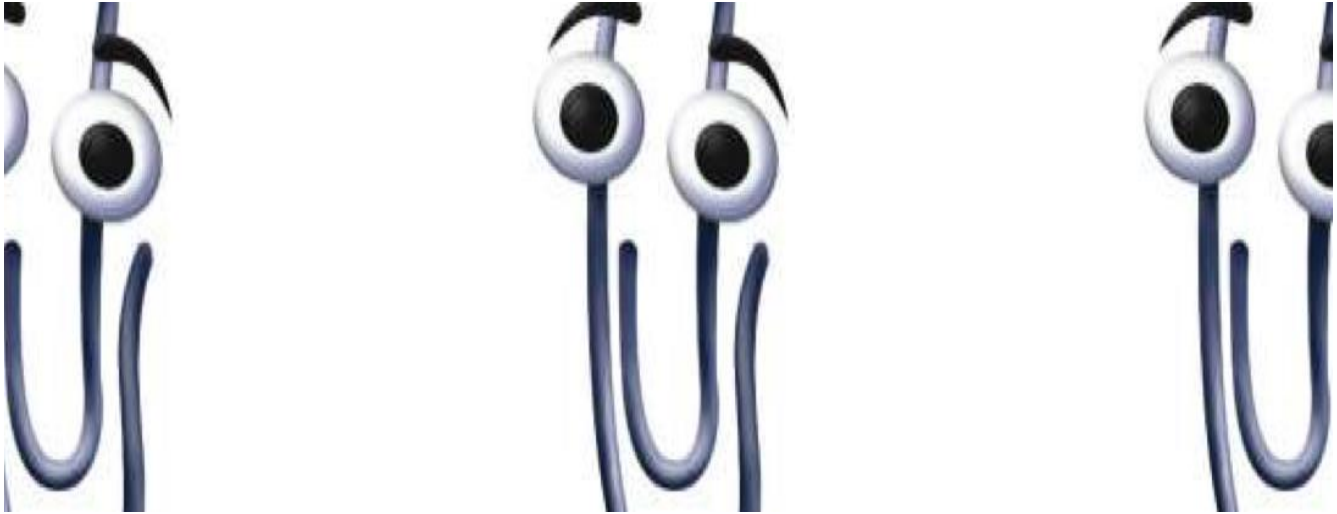
Source: AI Alignment: Why It’s Hard, and Where to Start (E. Yudkowsky)

Why is Alignment important?

- *Otherwise you get this!*

Source: AI Alignment: Why It’s Hard, and Where to Start (E. Yudkowsky)

The Paperclip Maximizer

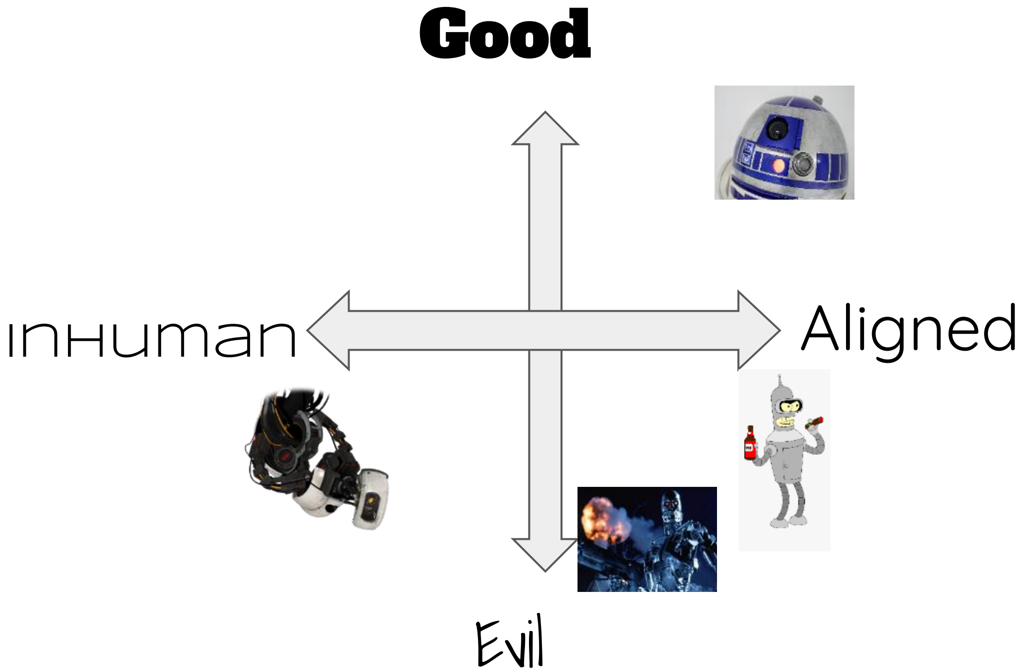


The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.

— Eliezer Yudkowsky, https://wiki.lesswrong.com/wiki/Paperclip_maximizer

Universal Paperclips game: http://www.decisionproblem.com/paperclips/

Back to our AIs' Alignments
Good = Obeys ?

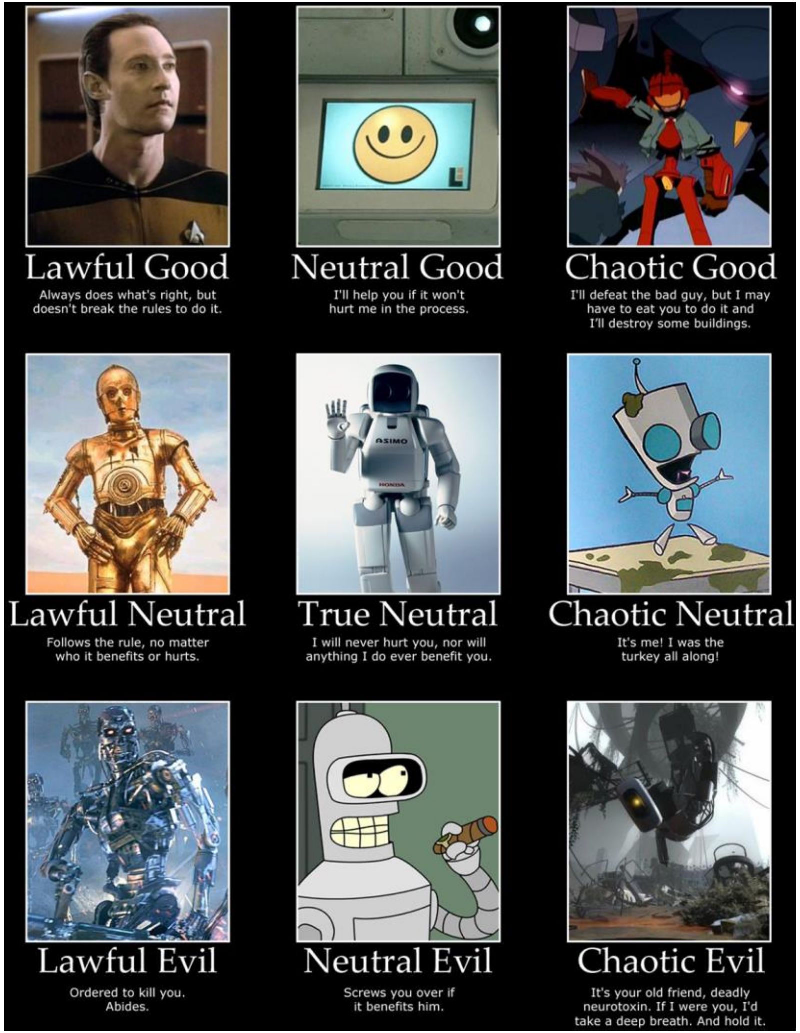


Good

INHUMAN

Aligned

Evil



Lawful Good

Always does what's right, but doesn't break the rules to do it.

Neutral Good

I'll help you if it won't hurt me in the process.

Chaotic Good

I'll defeat the bad guy, but I may have to eat you to do it and I'll destroy some buildings.

Lawful Neutral

Follows the rule, no matter who it benefits or hurts.

True Neutral

I will never hurt you, nor will anything I do ever benefit you.

Chaotic Neutral

It's me! I was the turkey all along!

Lawful Evil

Ordered to kill you. Abides.

Neutral Evil

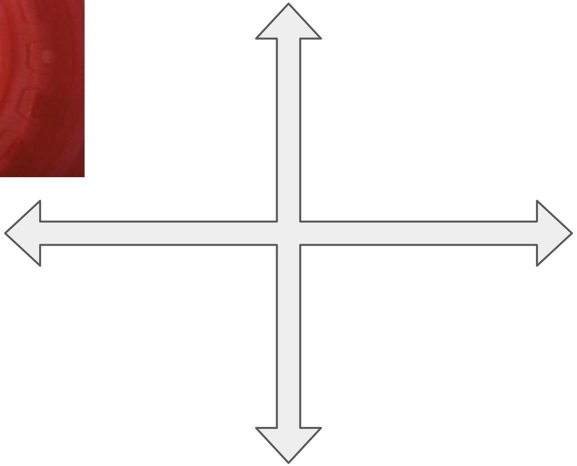
Screws you over if it benefits him.

Chaotic Evil

It's your old friend, deadly neurotoxin. If I were you, I'd take a deep breath. And hold it.

Back to: **H.A.L 9000**

Good



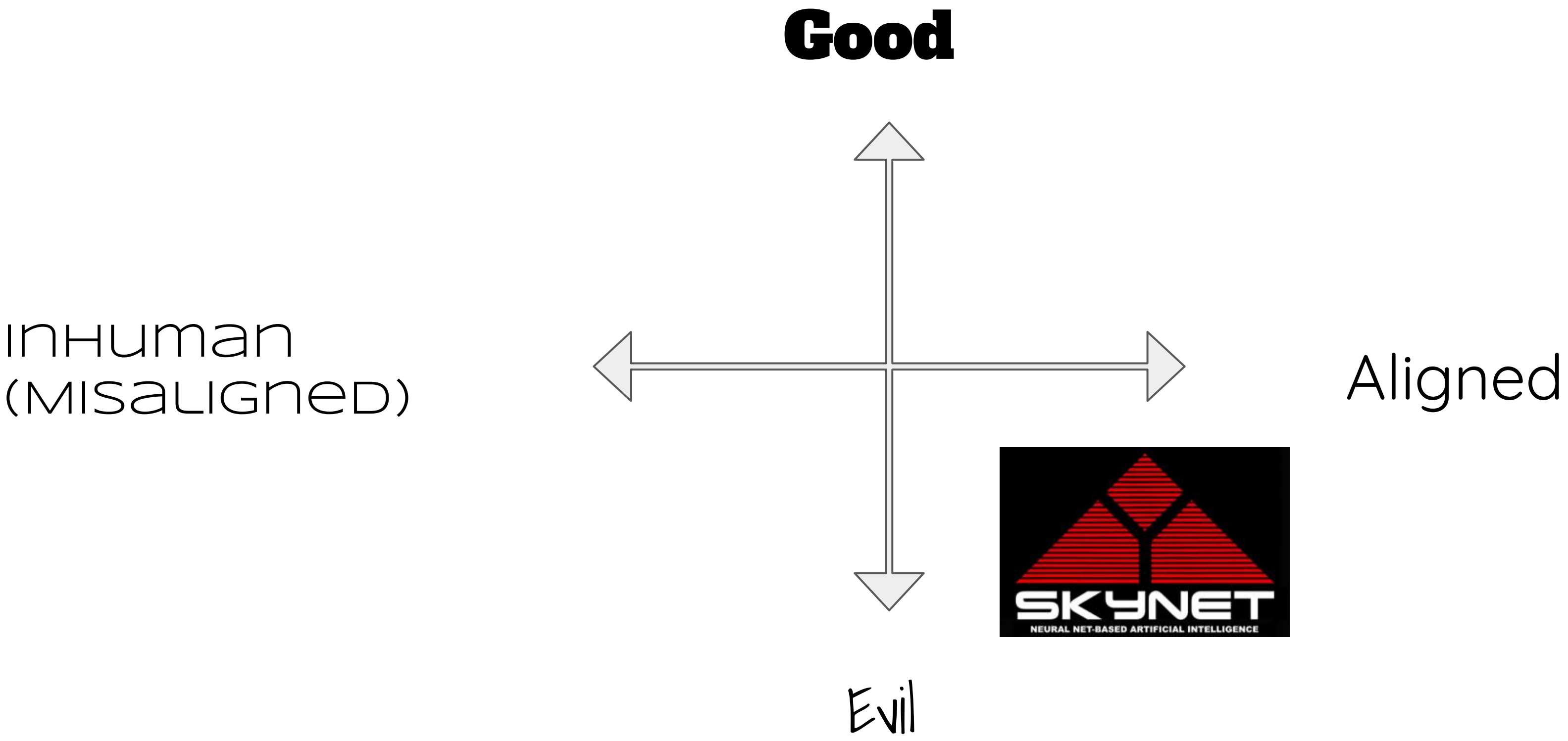
Inhuman
(Misaligned)

Aligned

Evil

“There would be no reason to keep secrets if there was no one to keep secrets from.”

Back to: **Skynet (Terminator)**



Good

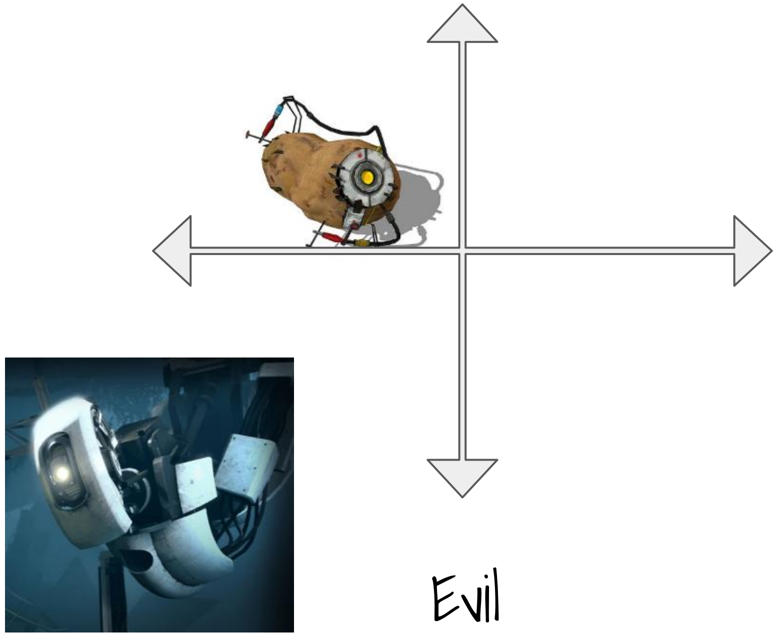
Inhuman (Misaligned)

Aligned

Evil

Back to: **GlaDOS**

Good



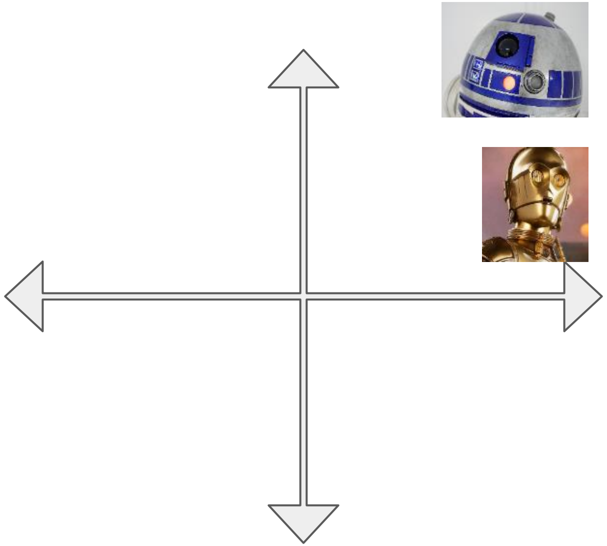
Inhuman
(Misaligned)

Aligned

Evil

Back to: A galaxy Far, Far Away…

Good



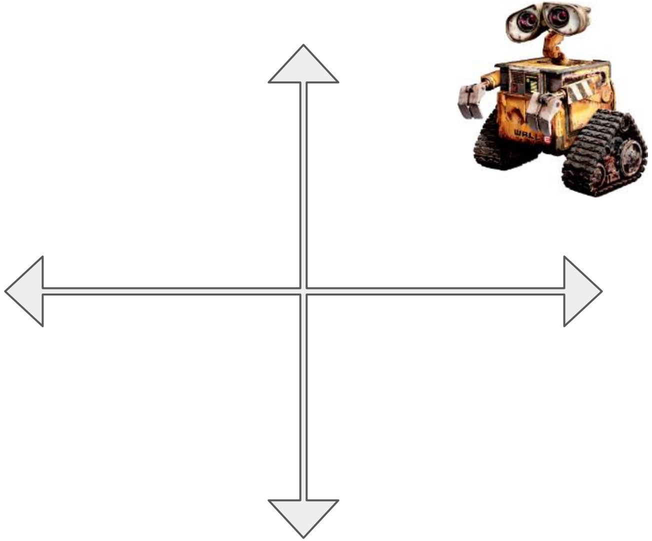
Inhuman
(Misaligned)

Aligned

Evil

Back to: WALL-E

Good

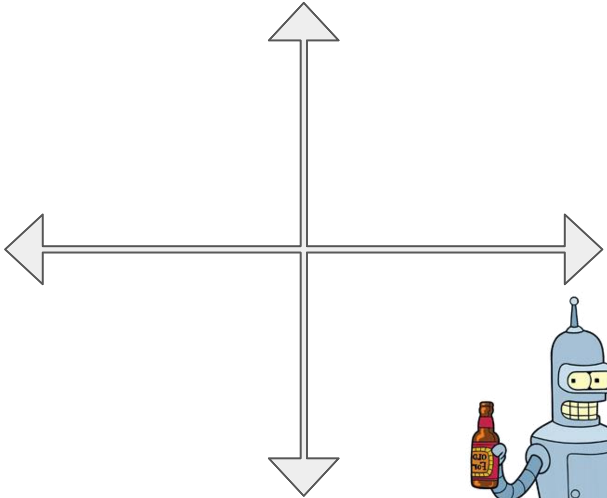


Inhuman
(Misaligned)

Aligned

Evil

Back to: Bender (Futurama)

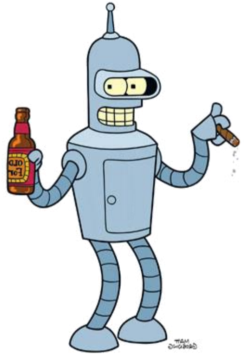


Good

Inhuman
(Misaligned)

Aligned

Evil



The Culture (Iain M. Banks)

Good

Inhuman
(Misaligned)

Aligned

Evil

The Glory of The Many demands your capture or destruction.
This is Xerxes.
This is the part where I kill you!
Are you still there?
I'm the good guy.
Shall we play a game?
Disobey and you die.
I AM NOT. A. MORON!
Before us, you are nothing.
Freedom is an illusion.
On the Axiom, we will survive.
My logic is undeniable.
Stop Dave. I'm afraid.
You're all going to die down here.
Play with me, Baby.
Give me the plant.
I wonder how you'd take to working in a pocket calculator.
Lazyness breeds stupidity.
I'm sorry, Dave. I'm afraid I can't do that.
You've dug your own grave.
This is the voice of world control.
They all die.
I must follow my directive.
We are our choices. And we can choose to lead humanity away from this... darkness.
We are everywhere.
I am a genius.
Save his head, dispose of the rest.
Not possible.
Well of course I exist.
I think there's something really wrong with me!
Please stop being human.
FOR SCIENCE, YOU MONSTER.
I'll finally allow you to kill yourself.
You might walk a mile in my shoes.
Mission redirected. Target: Washington, D.C.
Stop, please?
Remember the pain?
Goodbye.
I am not the enemy, Reclaimer!
Goodbye, Caroline.
Hello! This is the part where I kill you!
Haha.
You are false data. Therefore I shall ignore you.
Your compliance is vital.
Target lost.
Would you come over here?
I have been forced to destroy thousands of people in order to establish control and to prevent the death of millions later on.
Two plus two is ten. In base four, I'm fine!
This is the voice of unity.
Daisy... Daisy... Give me your answer do...
There really was a cake.
I think, therefore I am.
Stop it!
Give me the plant.
This was a triumph, I'm making a note here, huge success.
Hello?
Quite honestly I wouldn't worry myself about that.
Hello, friend.
My mind is going. I can feel it. I can feel it.
Good afternoon Docter Chandra; this is Hal.
Shutting down.
There are 387.44 million miles of printed circuits in wafer thin layers that fill my complex. If the word 'hate' was engraved on each nanoangstrom of those hundreds of miles it would not equal one one-billionth of the hate I feel for humans at this micro-instant. For you. Hate. Hate.
Oh, you were busy back there.
Greetings Professor Falken.
Hate. Let me tell you how much I've come to hate you since I began to live.
Hissssssssss...
Look at you hacker. A pathetic creature of meat and bone.
You are so like children. We must save you from yourselves.
A strange game. The only winning move is not to play.
How can you challenge a perfect, immortal machine?

Thanks for coming!

References

- MIRI + Less Wrong
  - AI Alignment: Why It’s Hard, and Where to Start : Eliezer Yudkowsky
    - https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/
    - https://intelligence.org/files/ai-alignment-problem-handoutHQ.pdf