The Ethics of AI

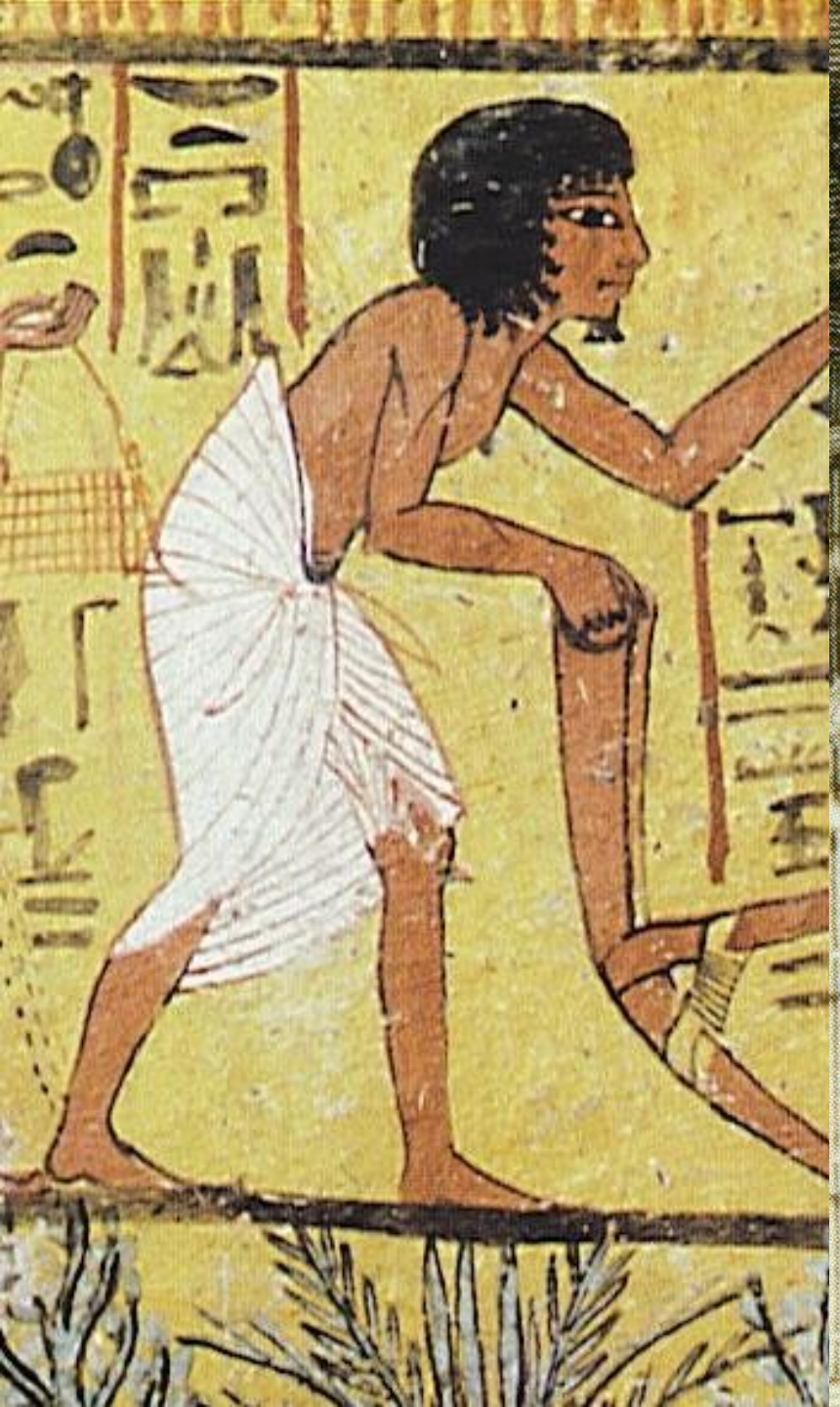
With Great Power Comes Great Responsibility

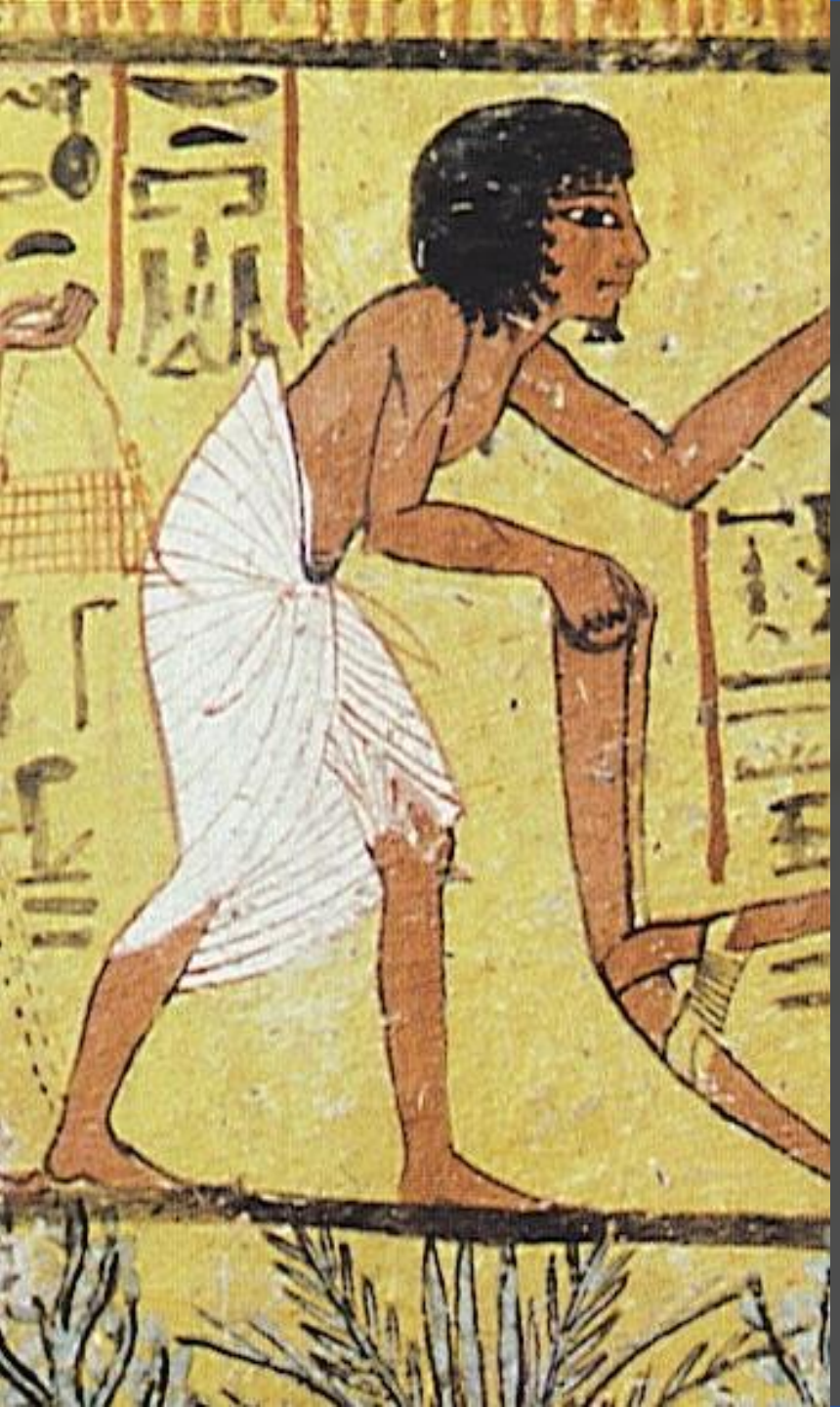
Matthew Renze

Renze Consulting

“With great power comes
great responsibility.”

-- Uncle Ben





Agricultural Superpower

Large armies

Abuses

Acquire land

Enslave labor

Subjugation

Industrial Superpower

Mechanized warfare

Abuses

Control resources

Maintain colonies

Expand influence

Information Superpower

Cyber warfare

Abuses

Mass surveillance
Industrial sabotage
Political propaganda

Purpose

To learn how to use AI responsibly and ethically

Overview

1. AI Ethics for Today

Overview

1. AI Ethics for Today
2. AI Ethics for Tomorrow

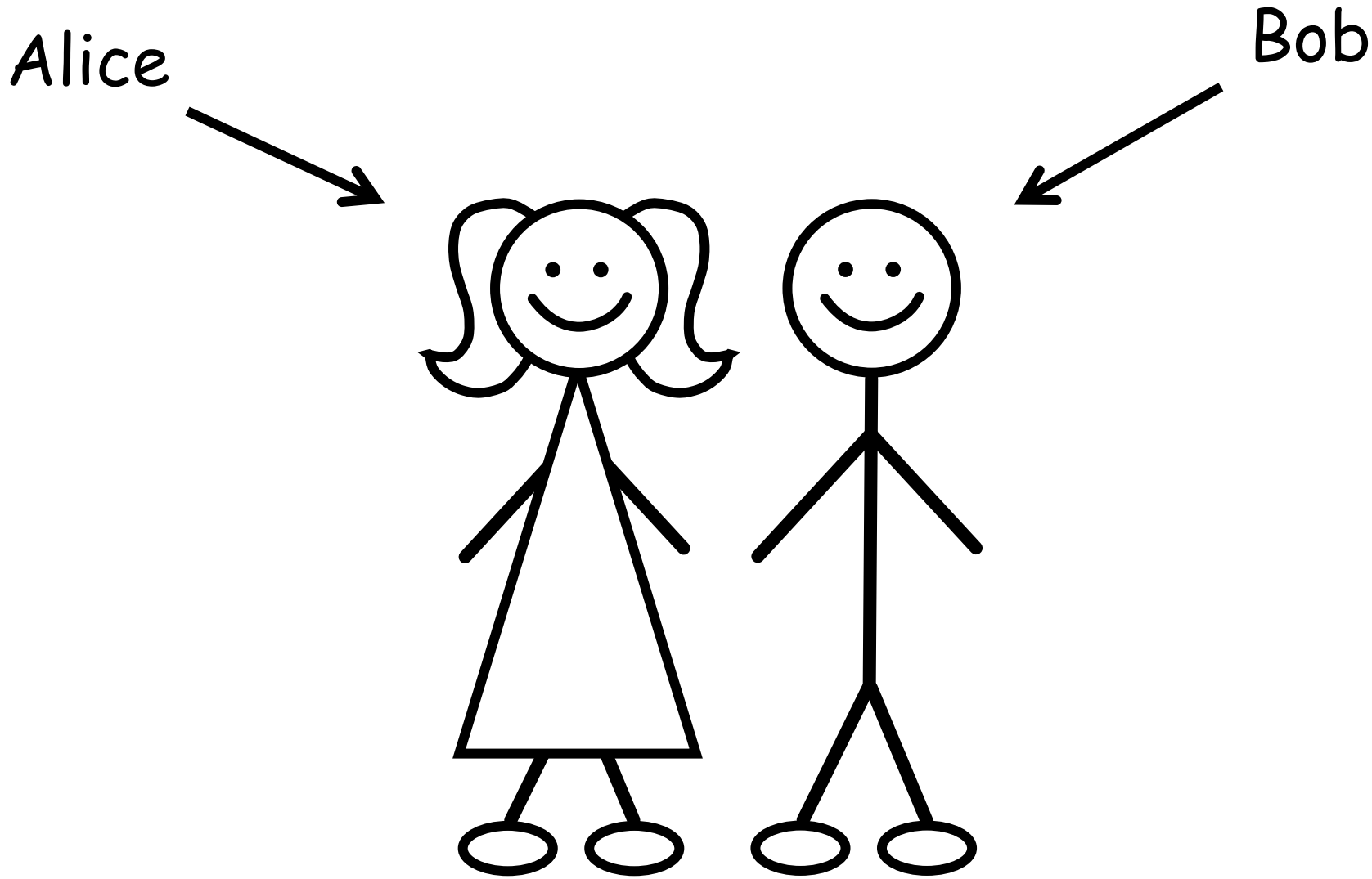
Overview

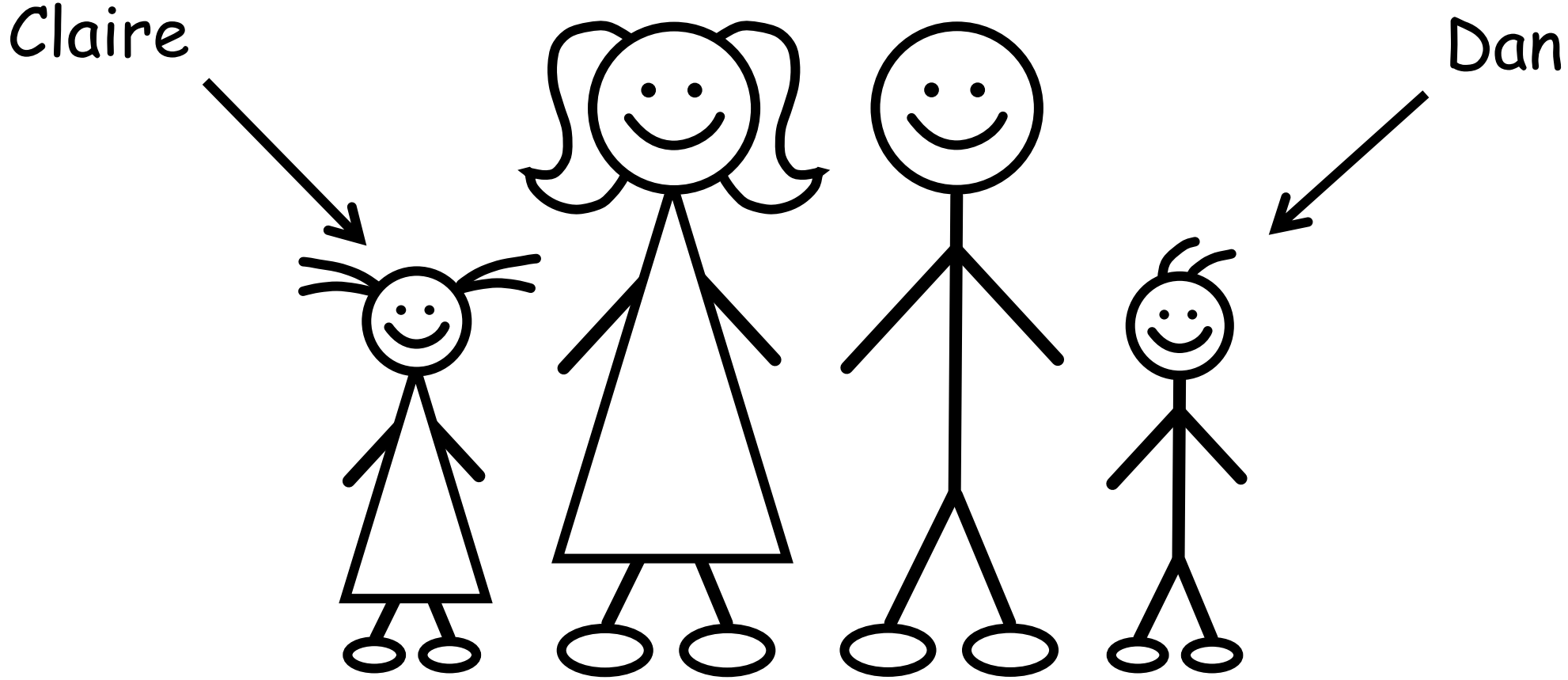
1. AI Ethics for Today
2. AI Ethics for Tomorrow
3. AI Ethics for The Future

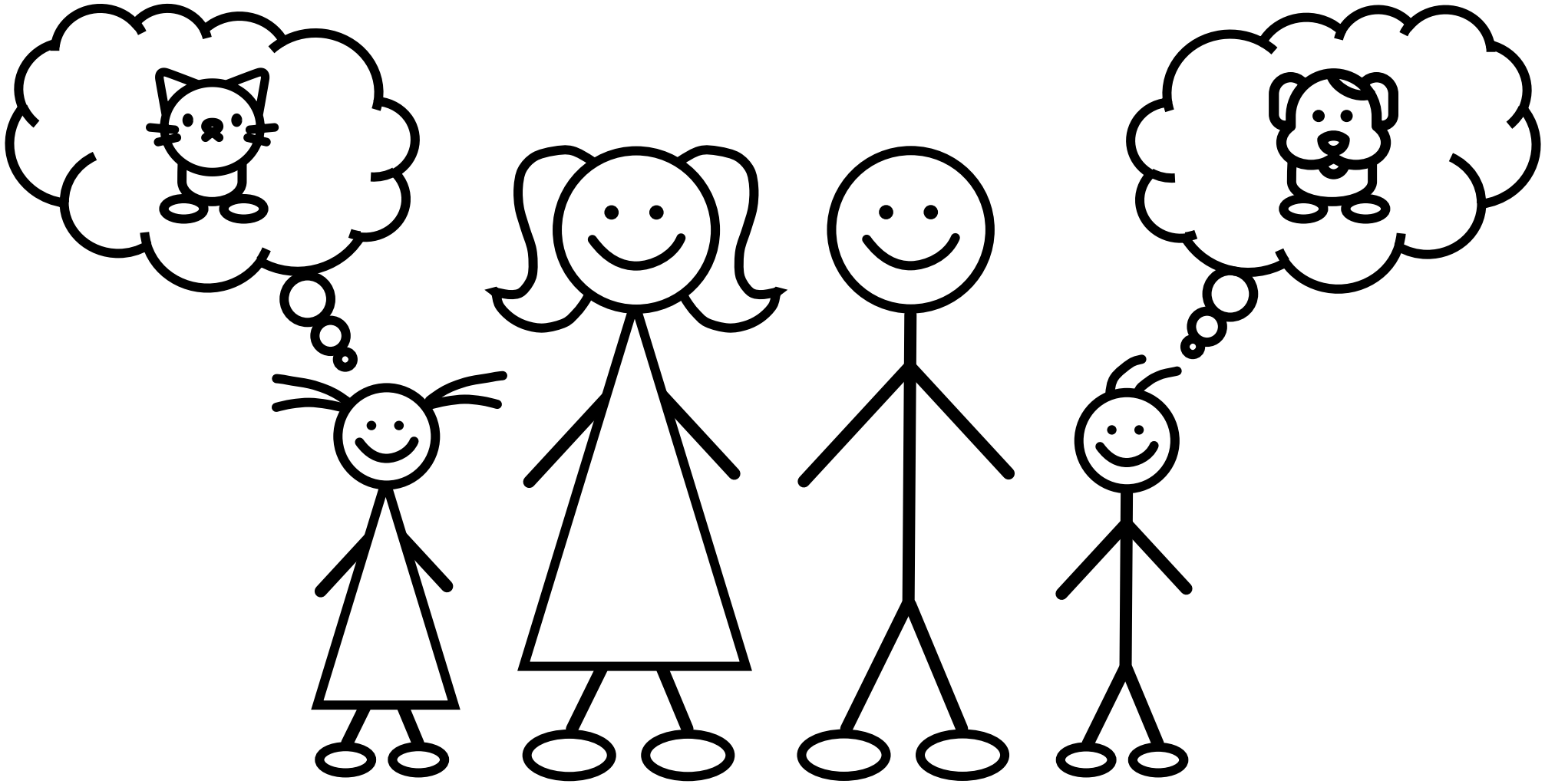
AI Ethics for Today

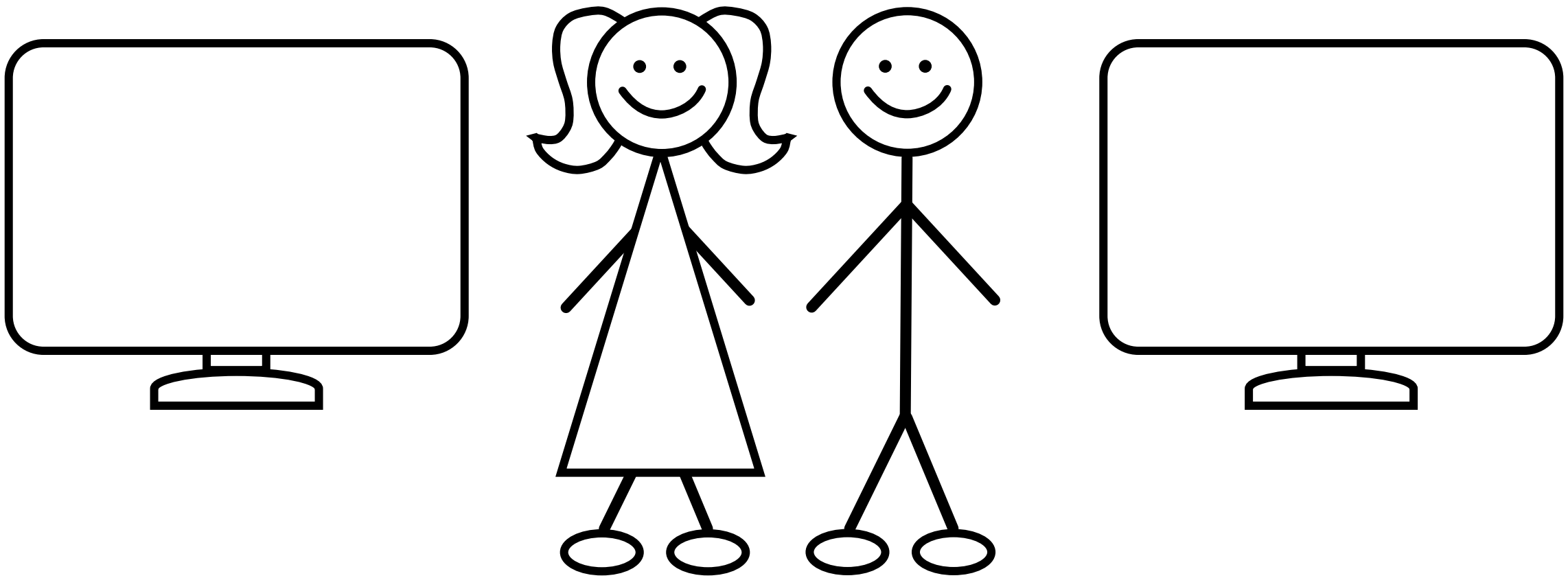
Alice and Bob in AI Land

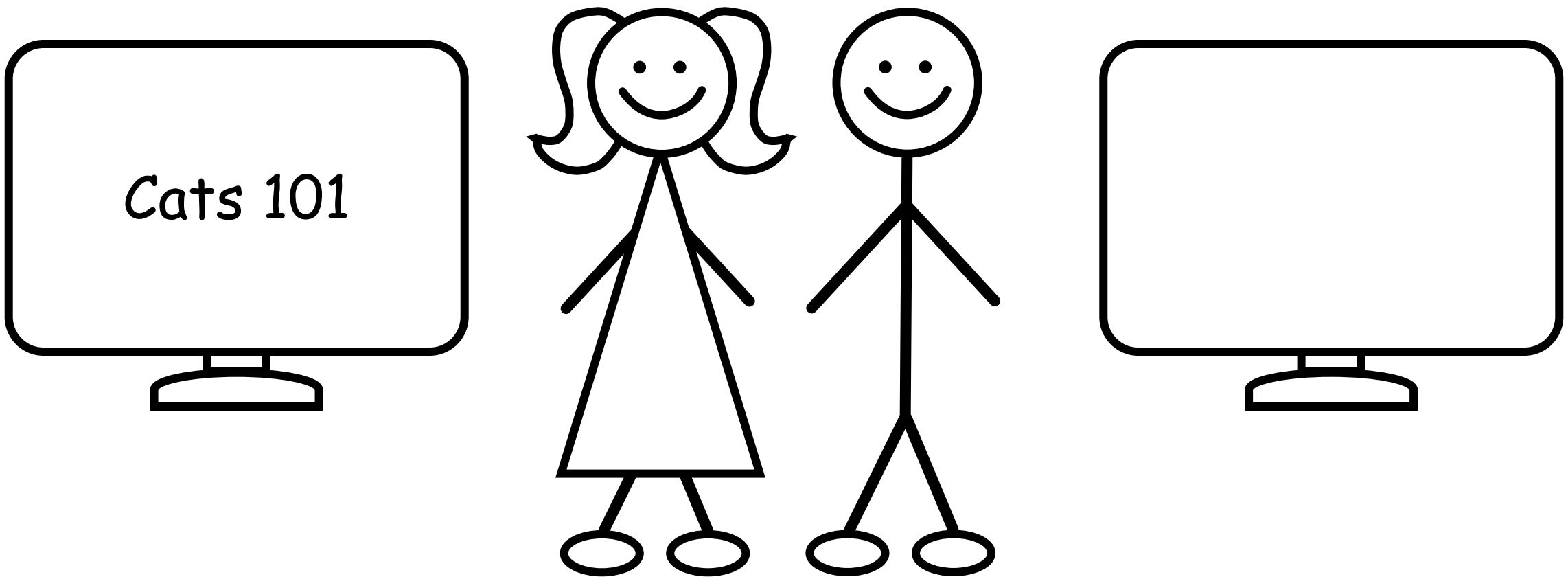
Chapter 1:
"AI Ethics of Today"

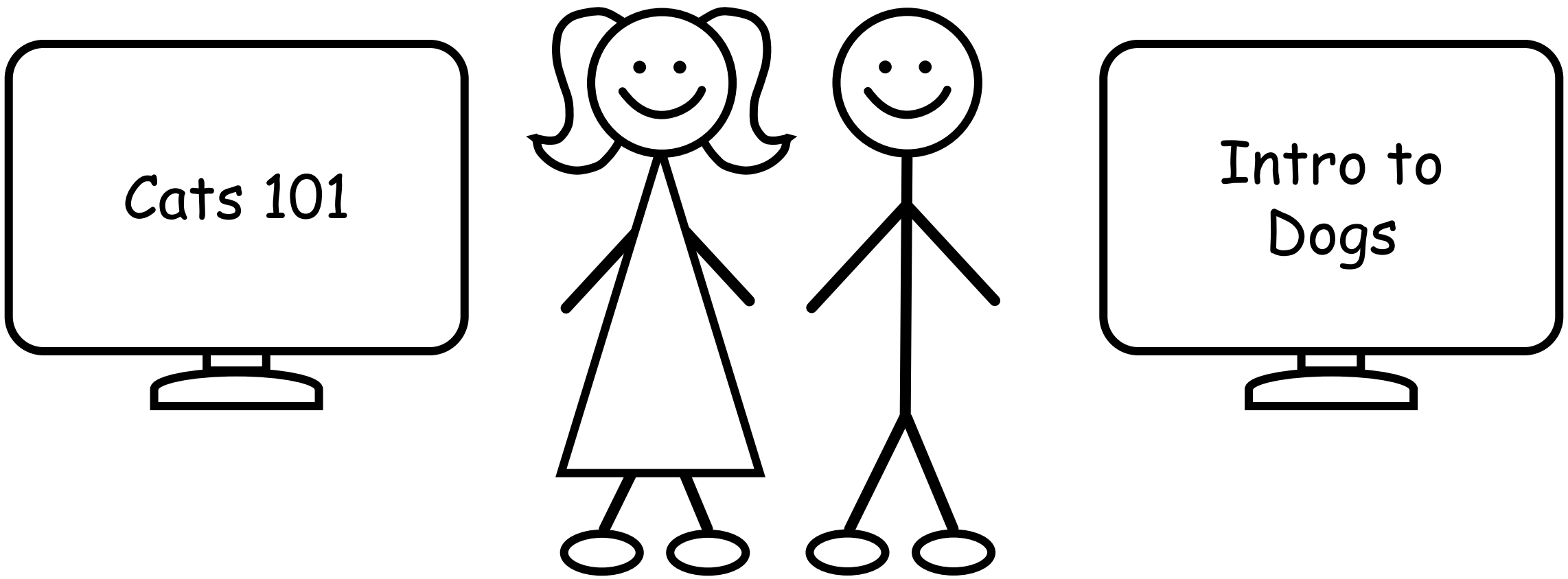

Alice
Bob

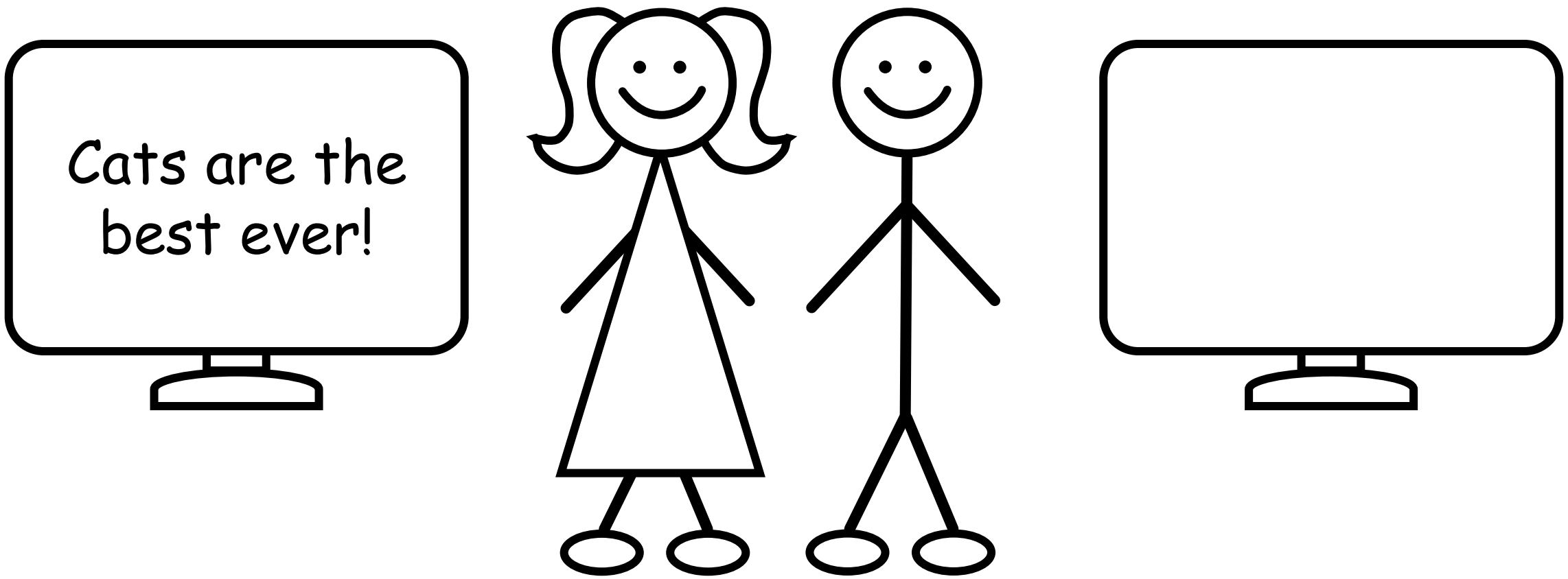

Claire
Dan

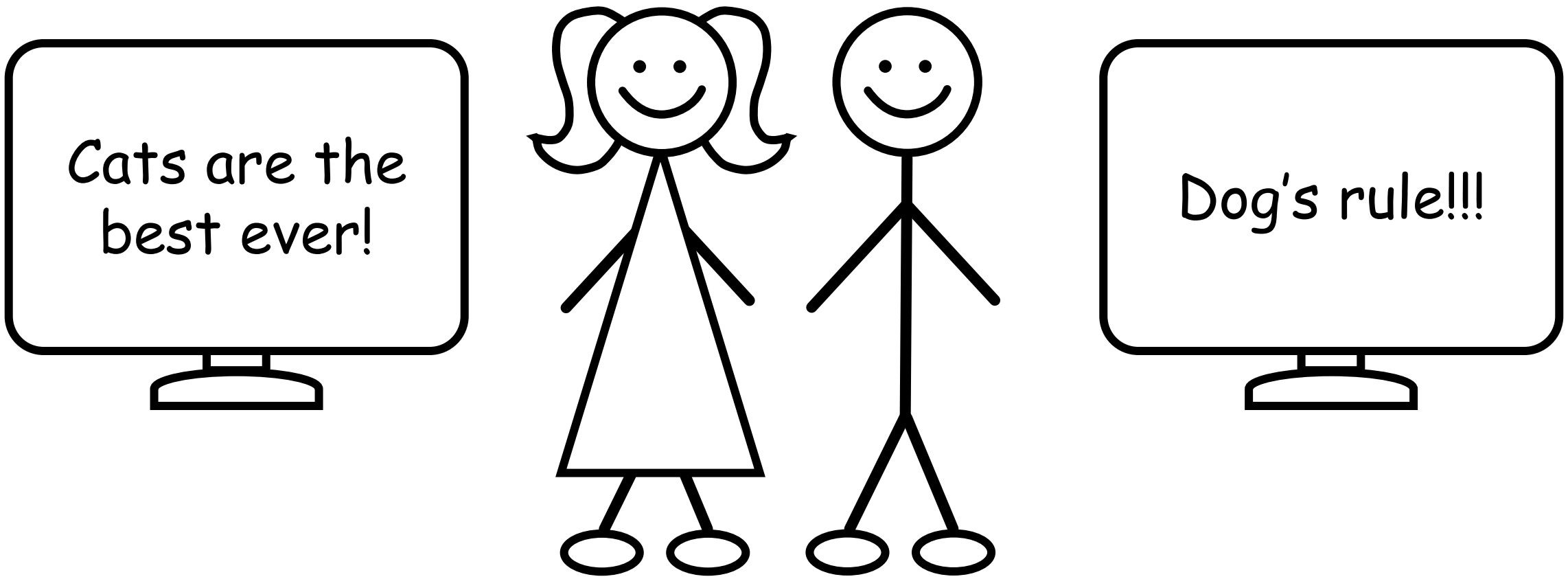


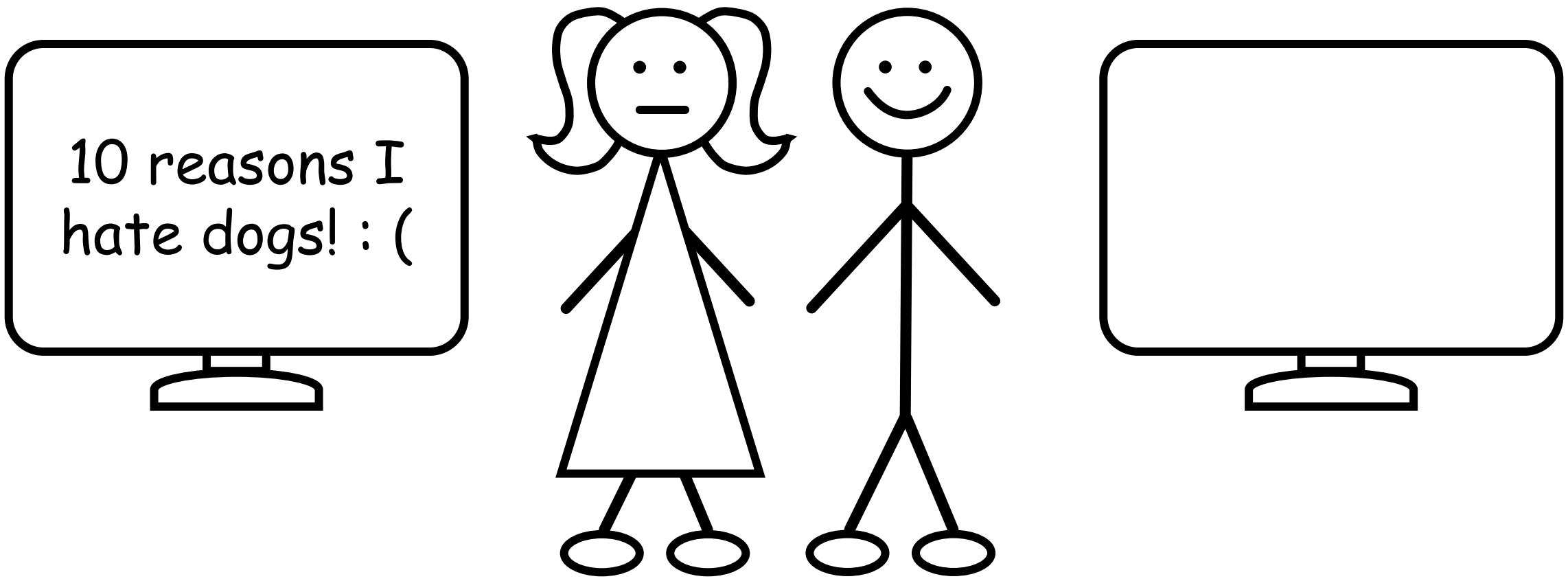


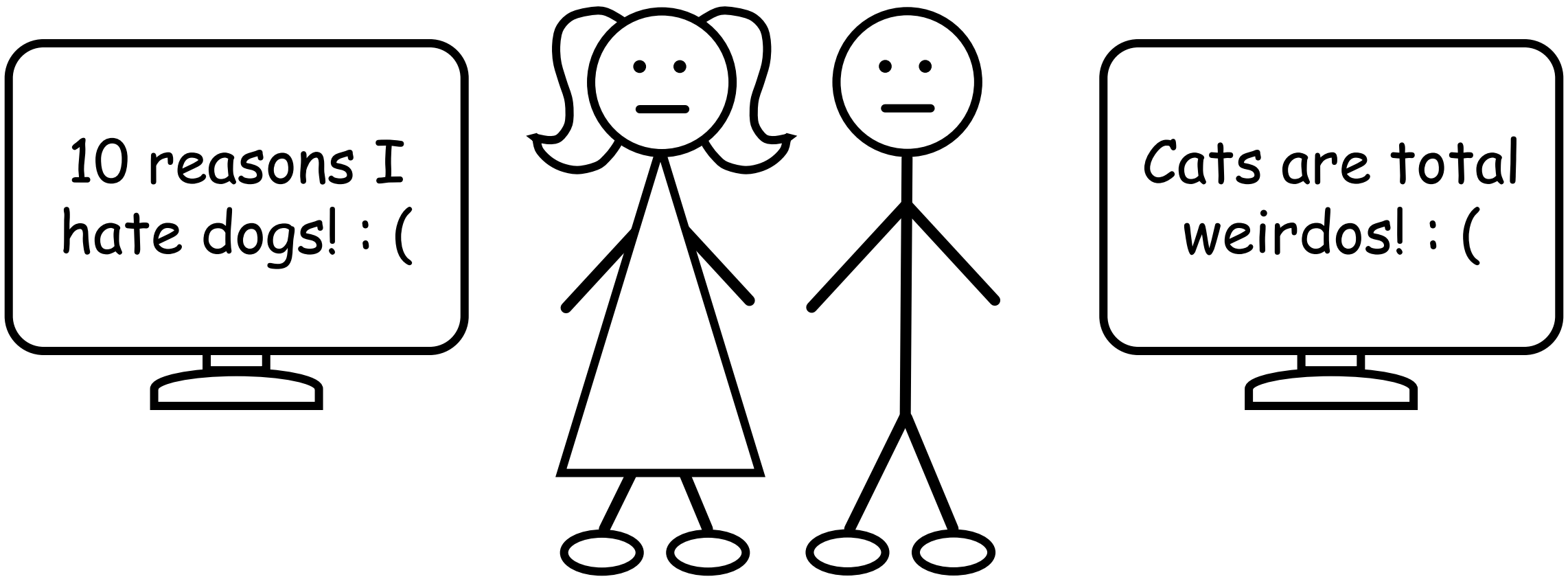

Cats 101

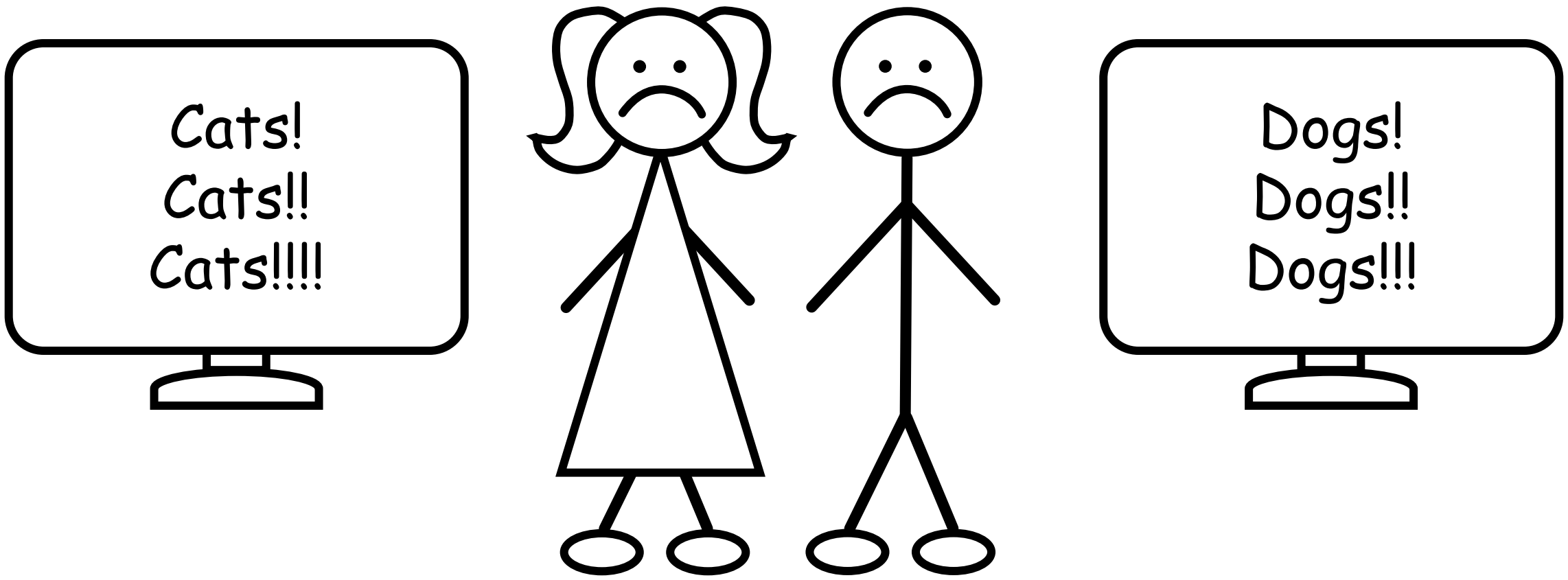

Cats 101
Intro to Dogs

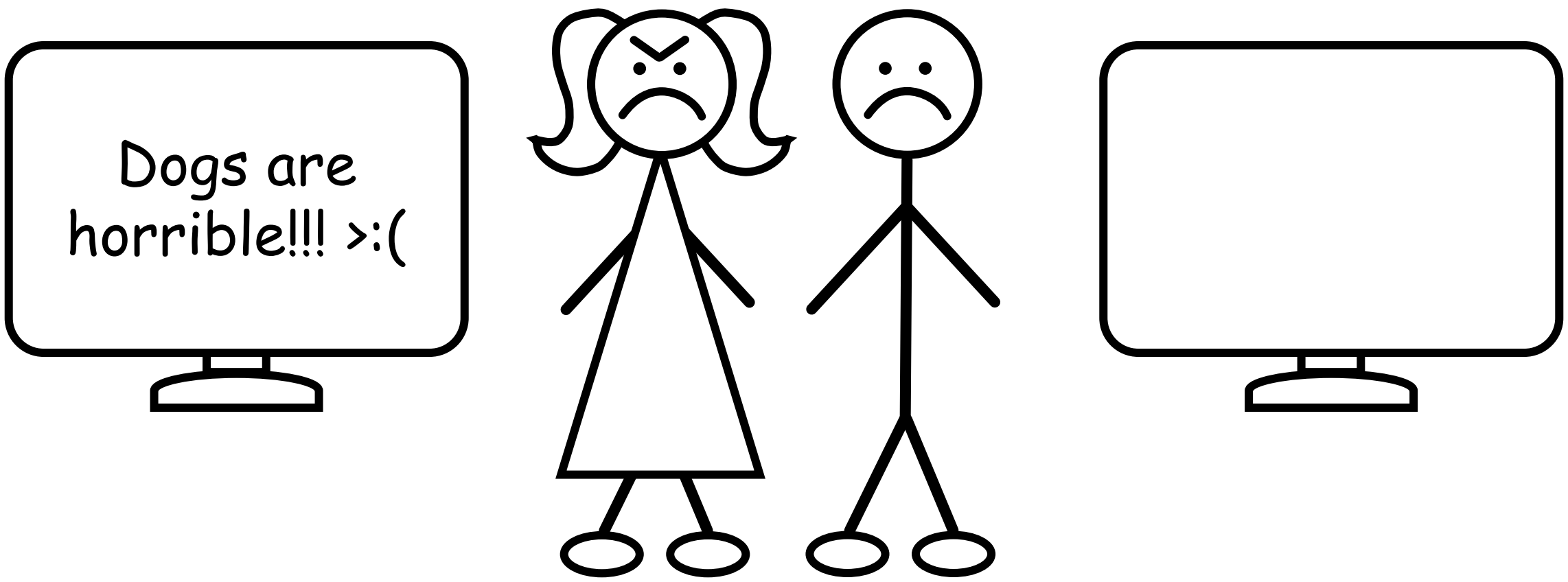

Cats are the best ever!

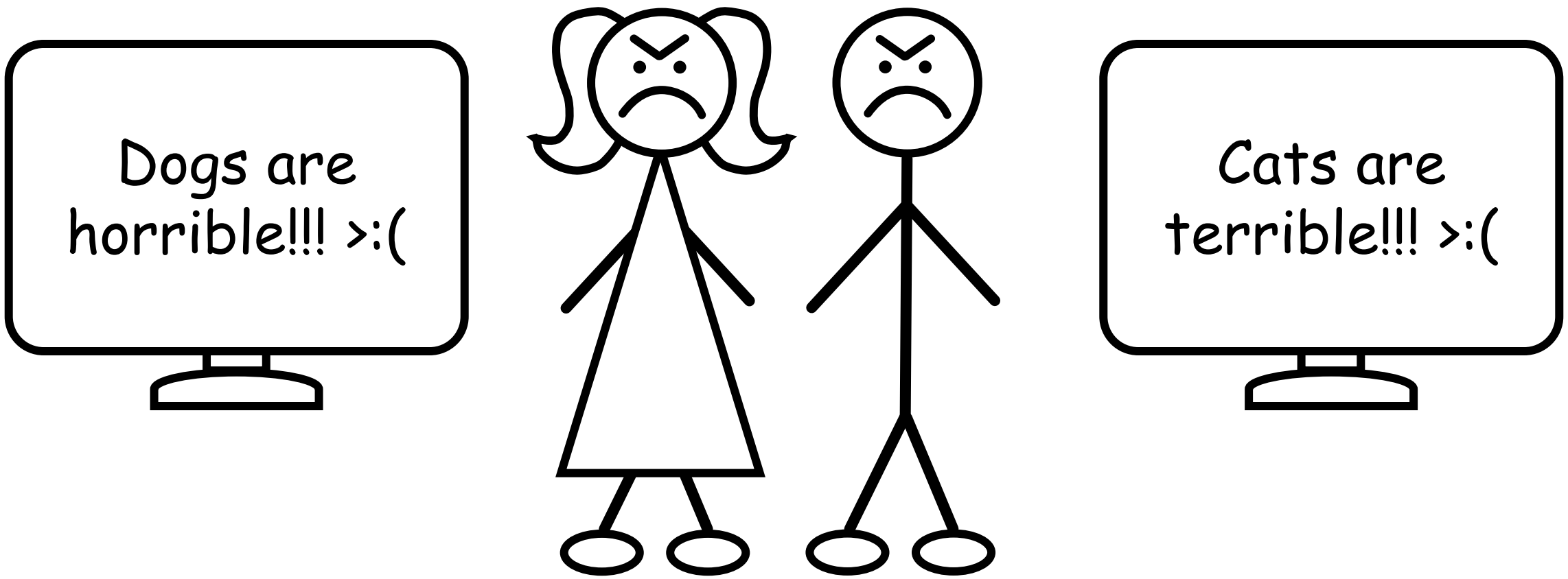

Cats are the best ever!
Dog's rule!!!

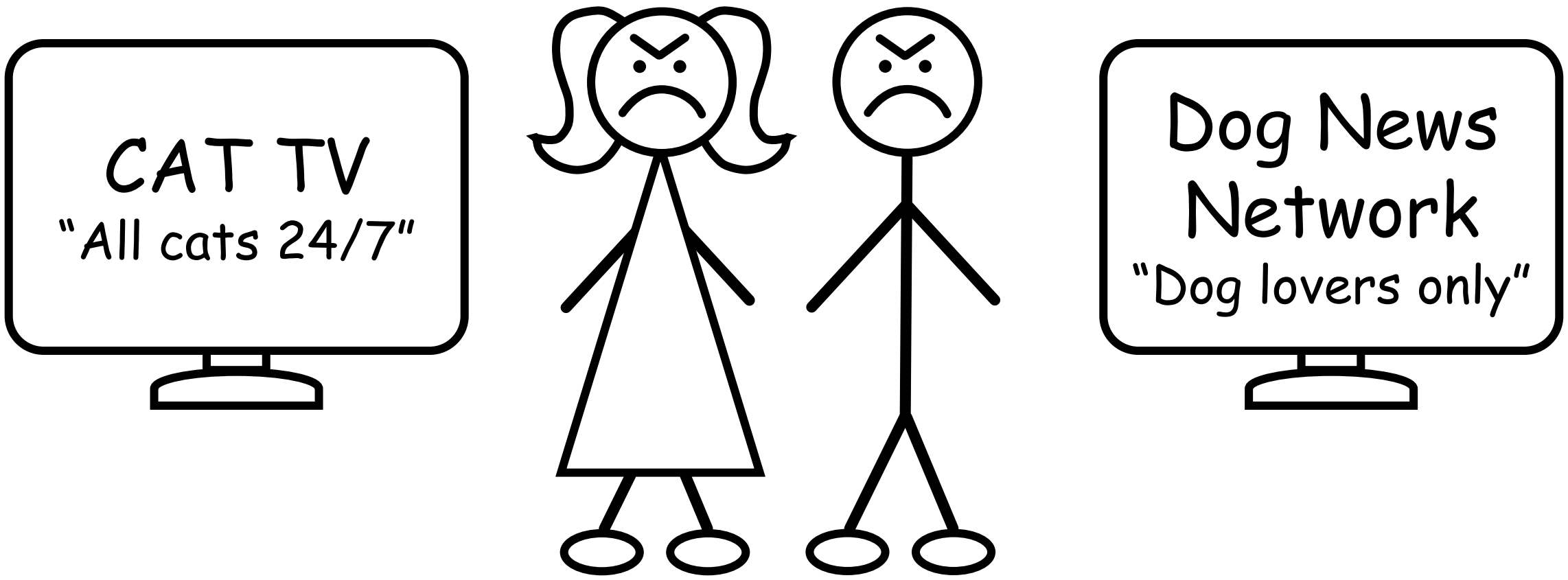

10 reasons I
hate dogs! : (

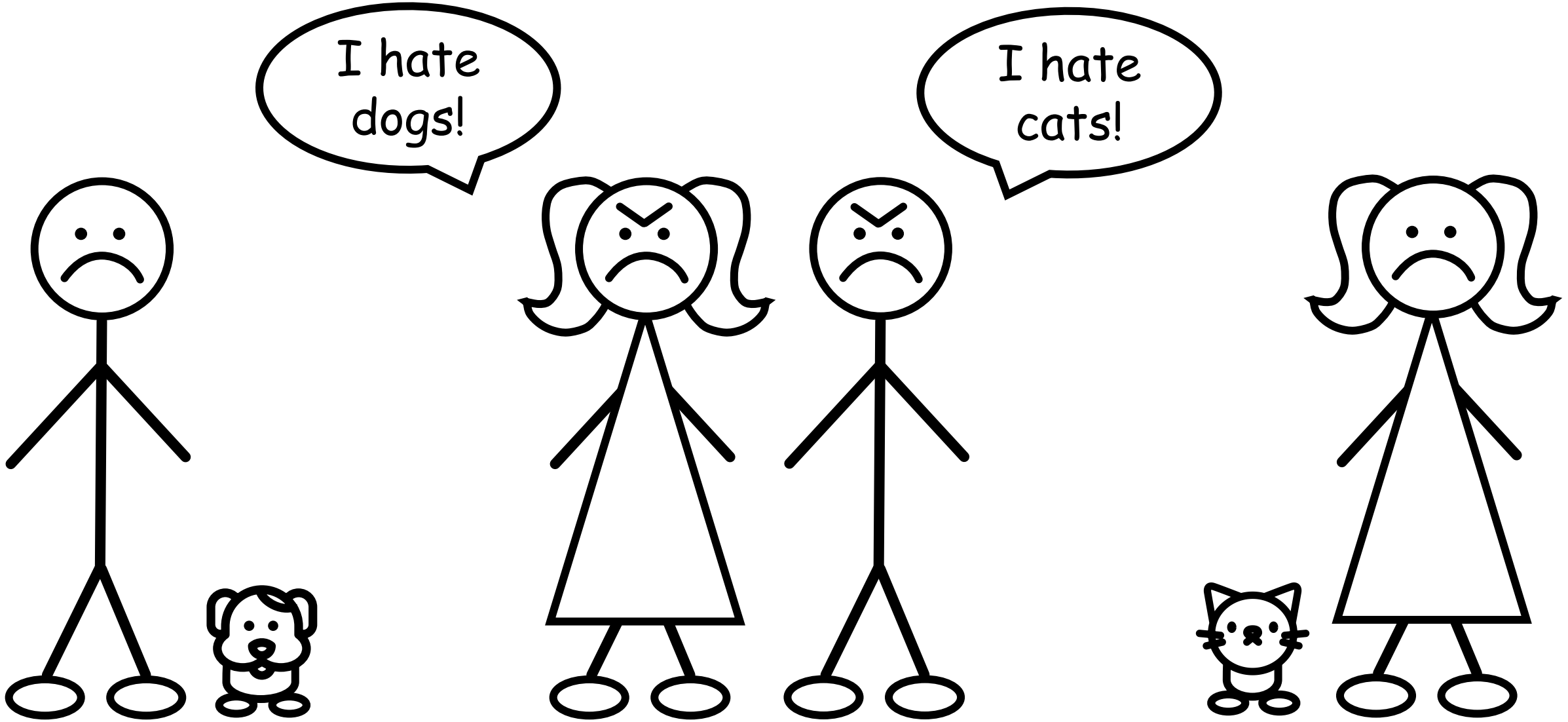

10 reasons I hate dogs! : (
Cats are total weirdos! : (

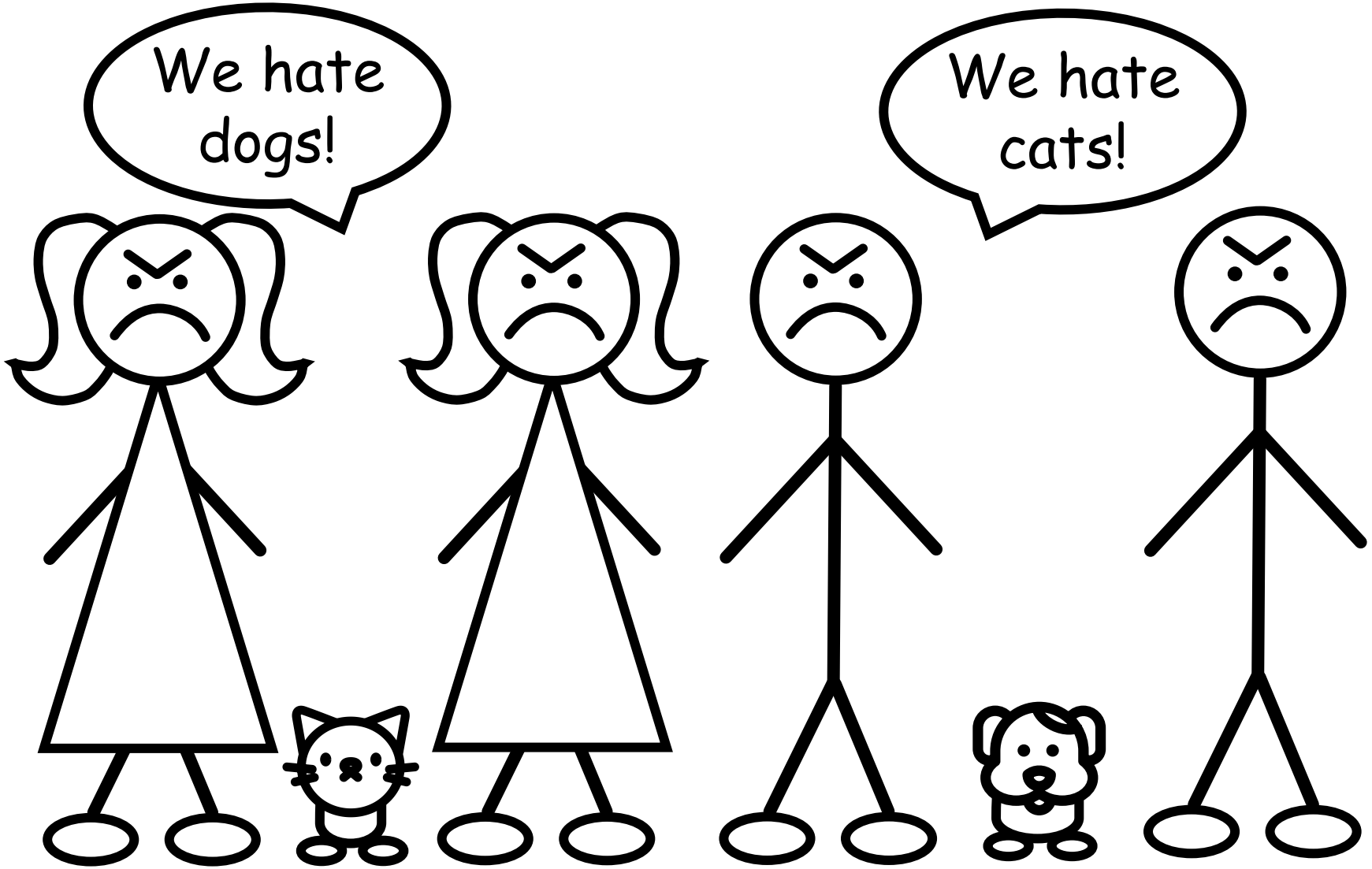

Cats!
Cats!!
Cats!!!!
Dogs!
Dogs!!
Dogs!!!

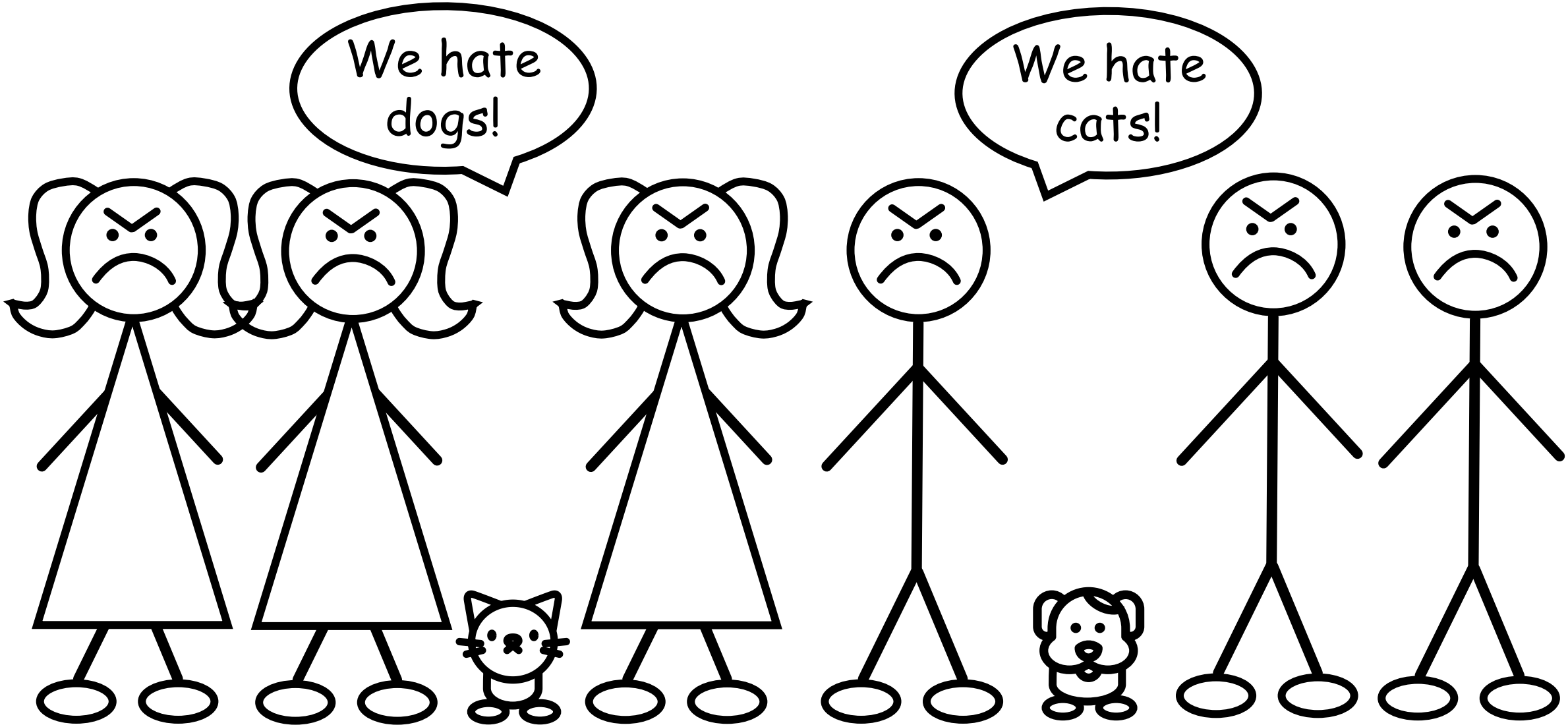

Dogs are horrible!!! >:(

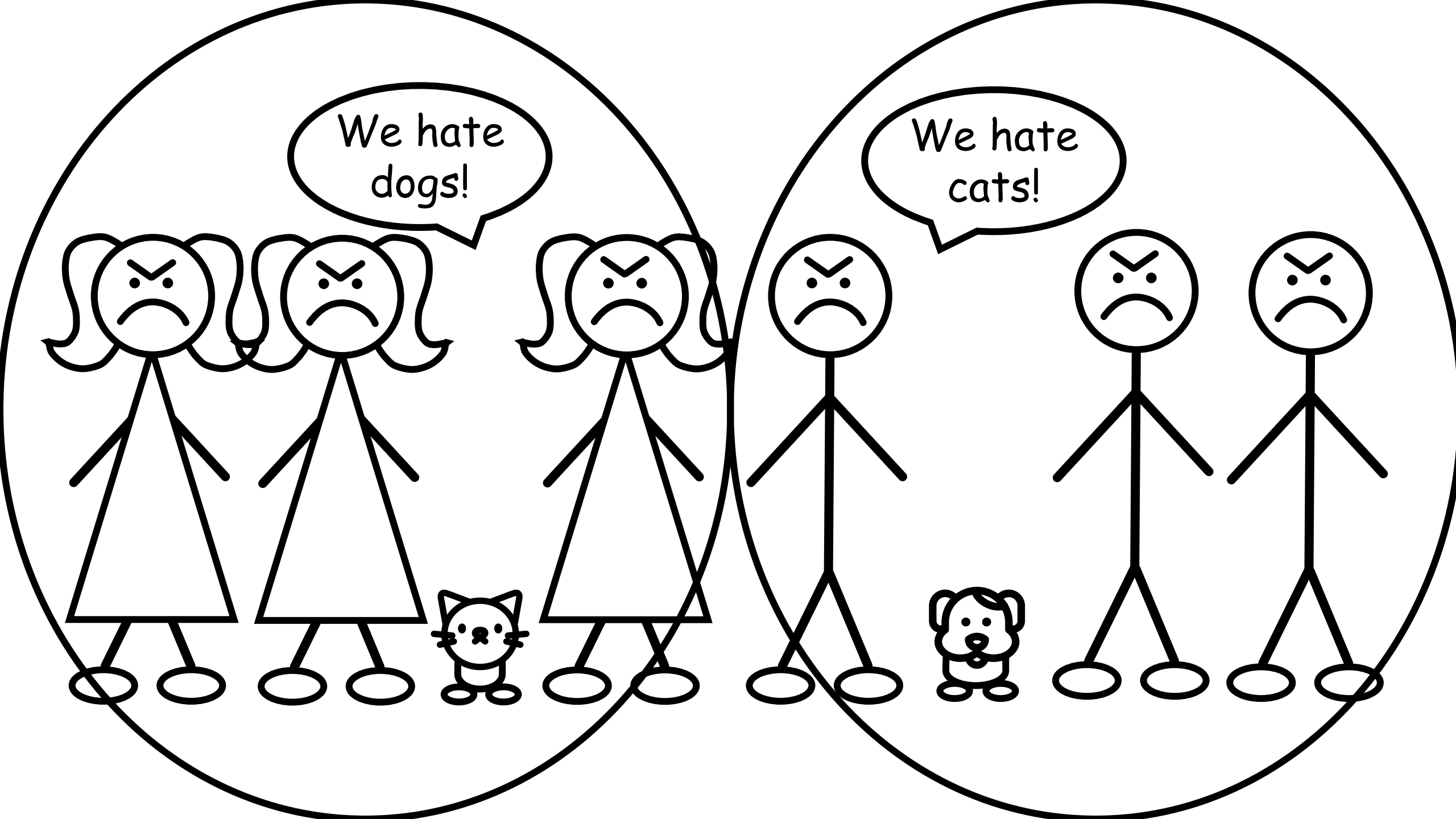

Dogs are horrible!!! >:(
Cats are terrible!!! >:(

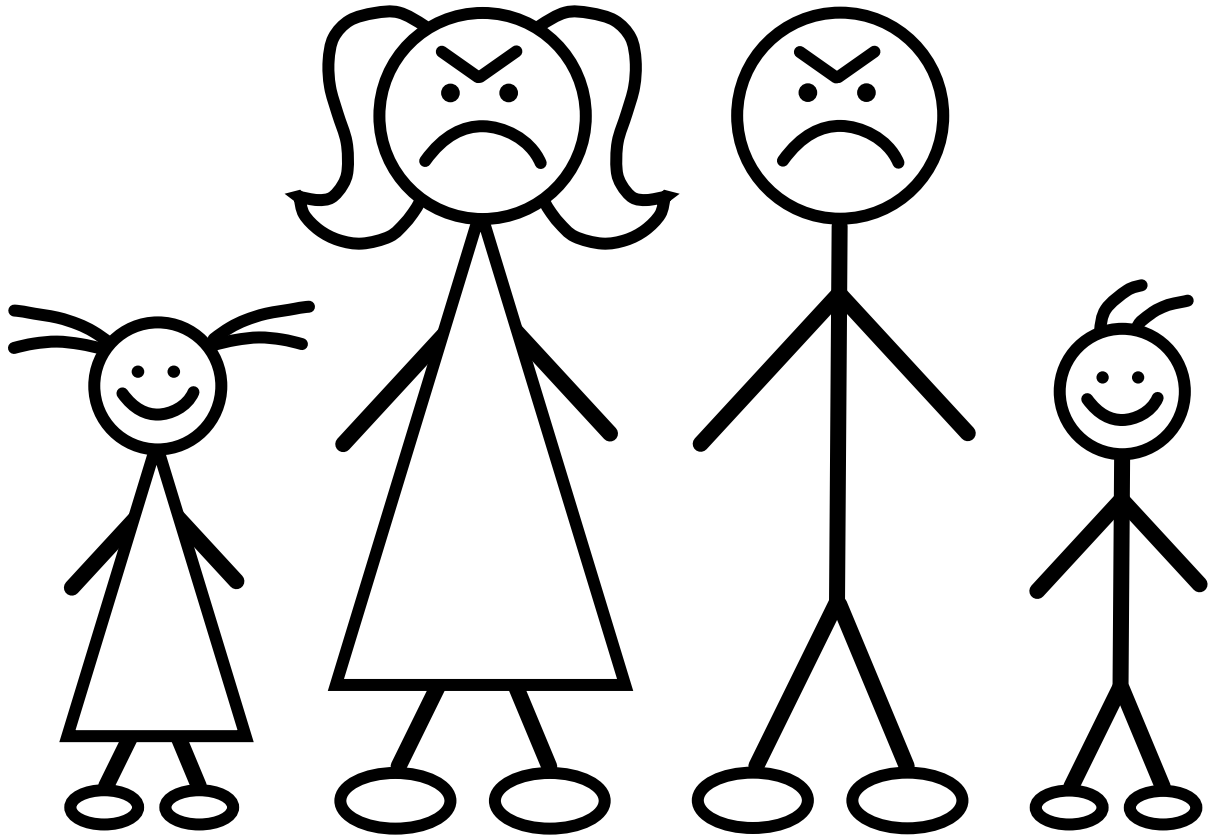

CAT TV
"All cats 24/7"
Dog News Network
"Dog lovers only"

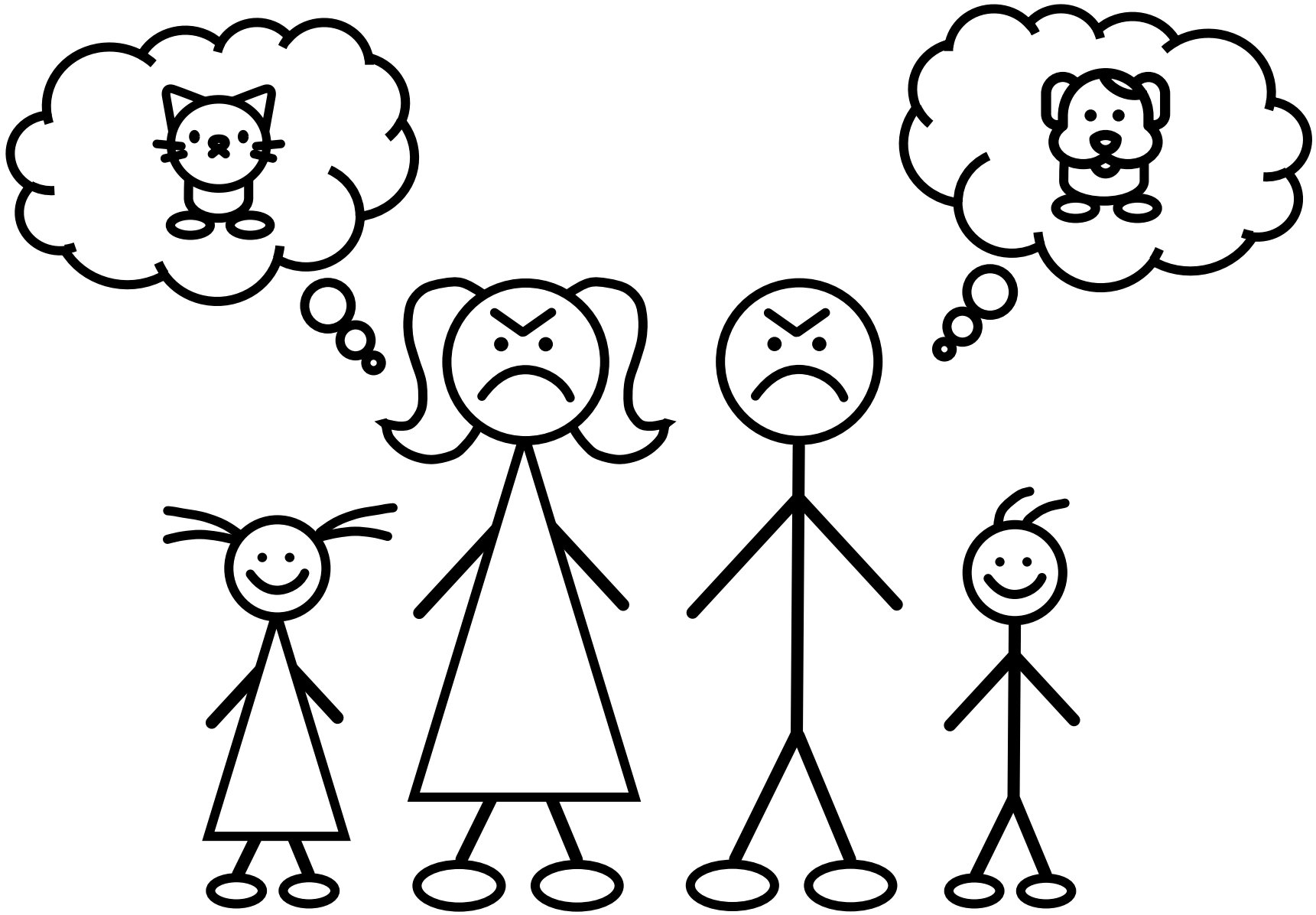

I hate dogs!
I hate cats!

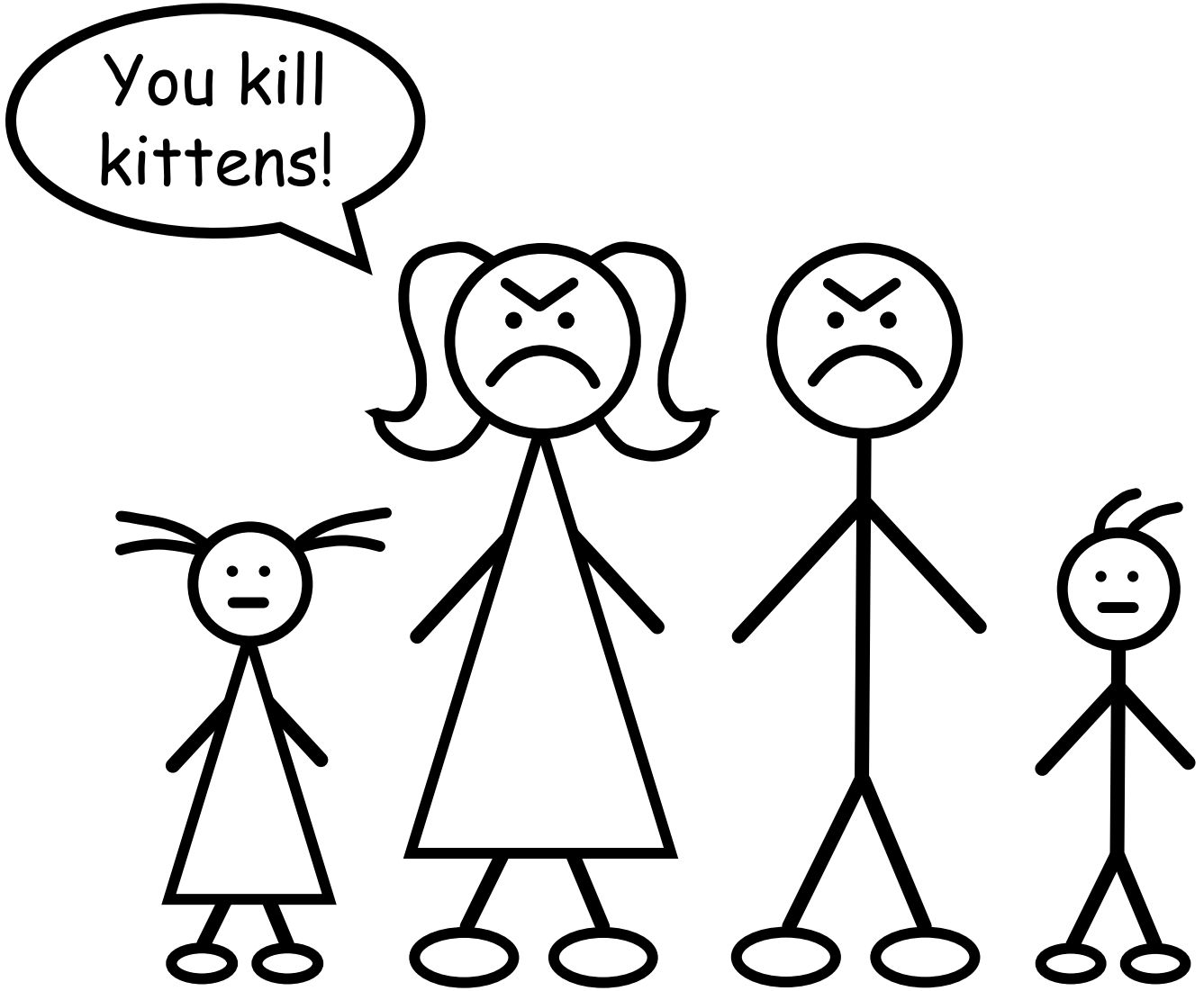

We hate dogs!
We hate cats!

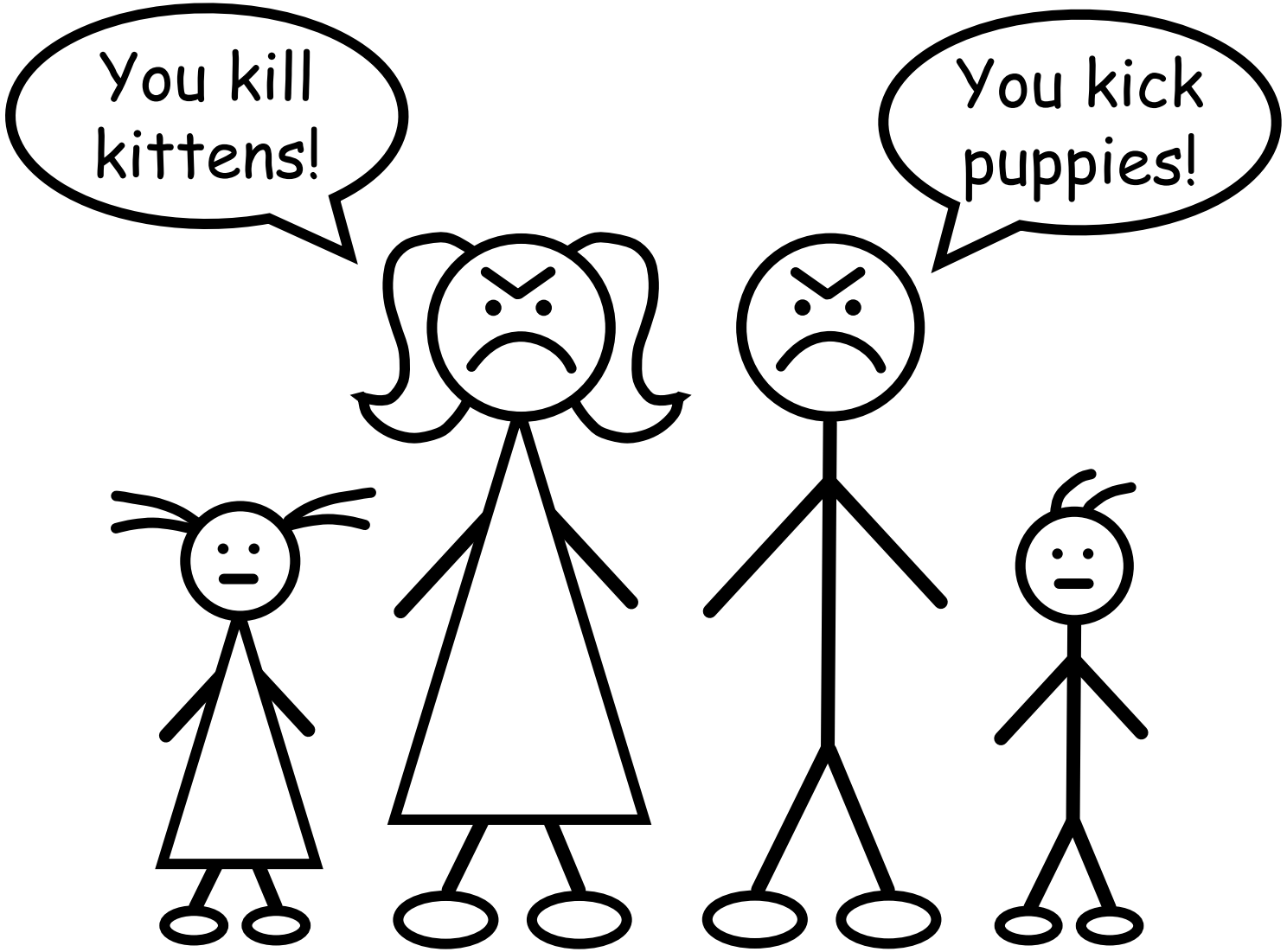

We hate dogs!
We hate cats!

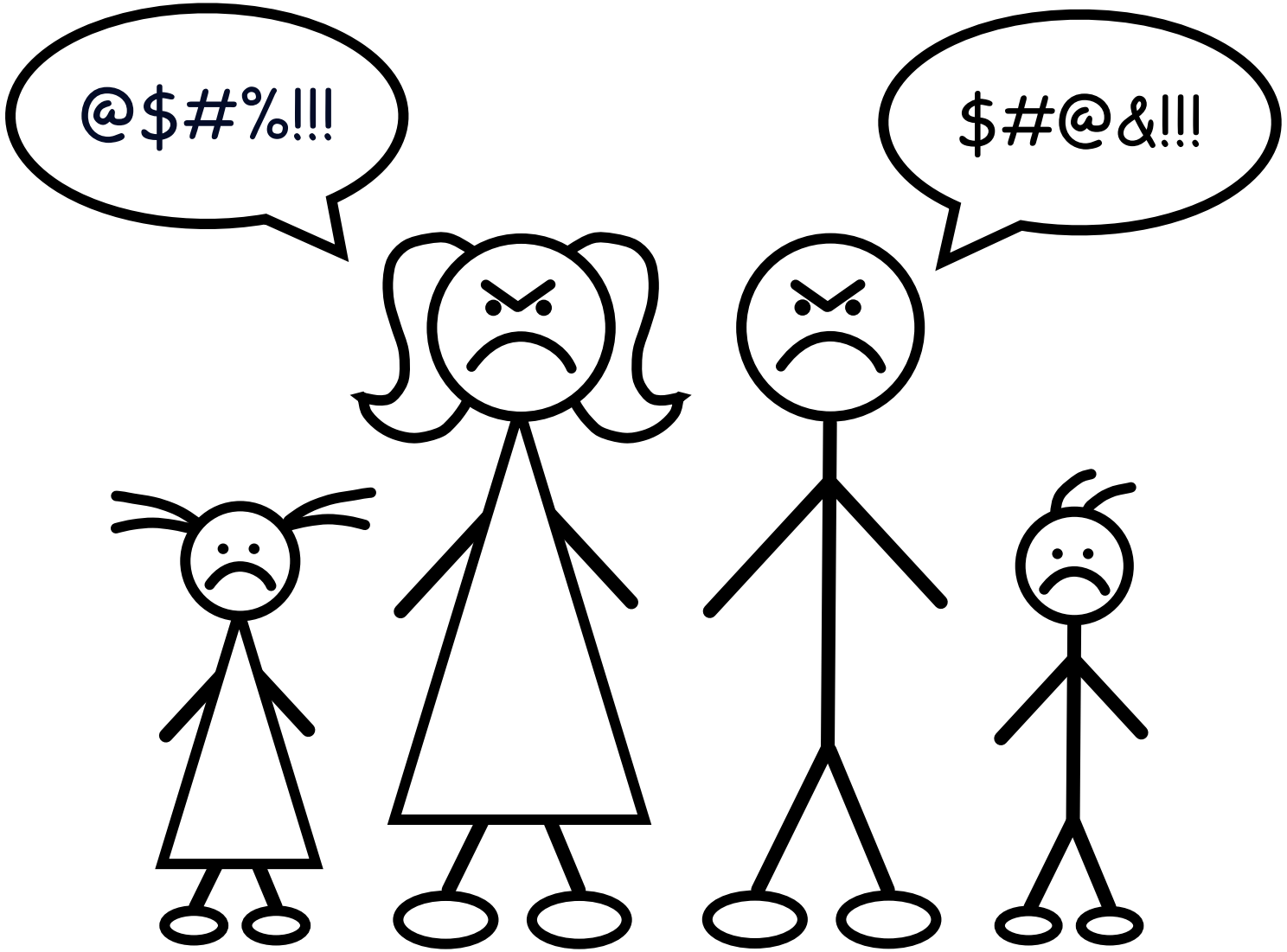

We hate dogs!
We hate cats!



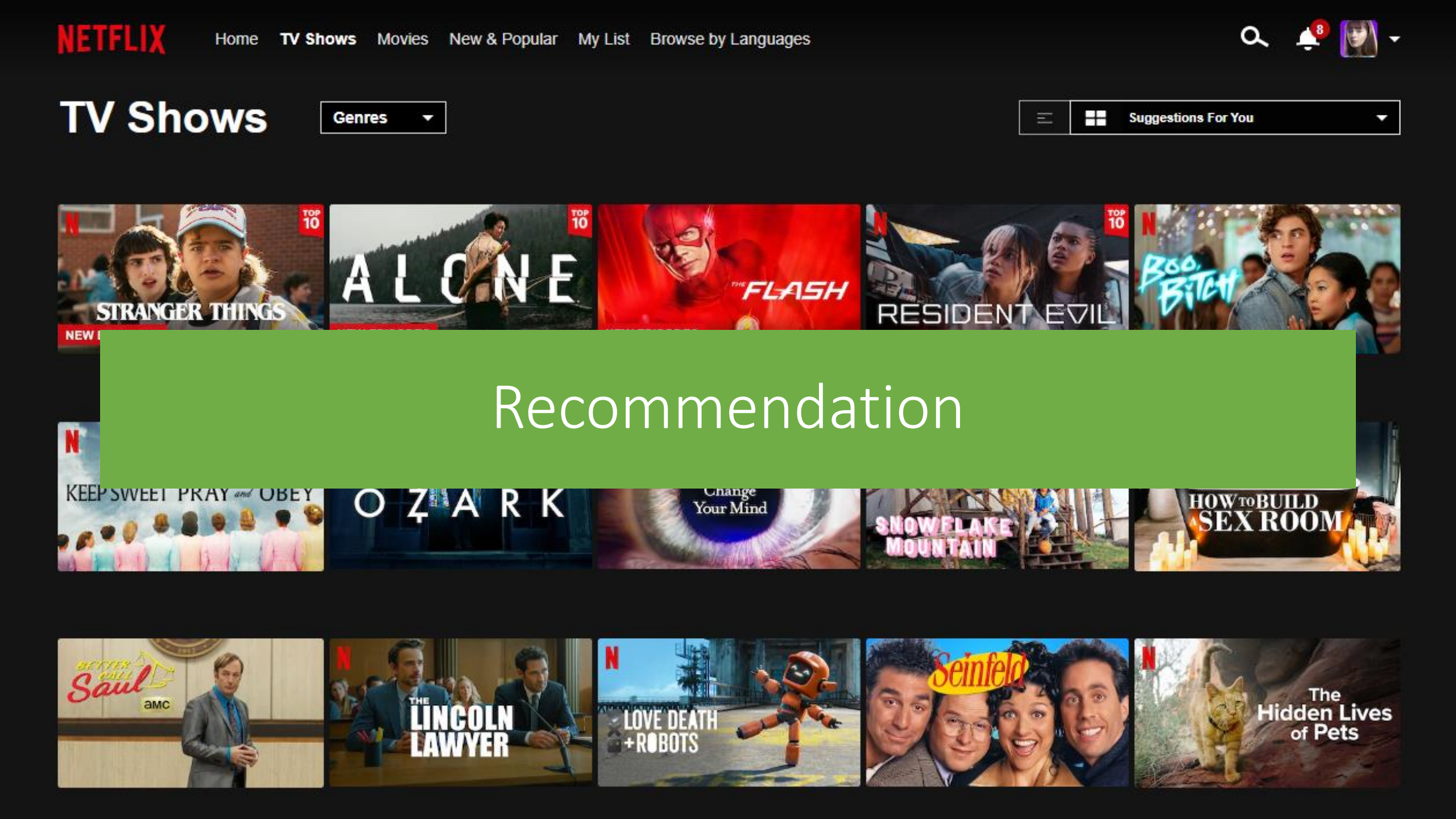



You kill kittens!

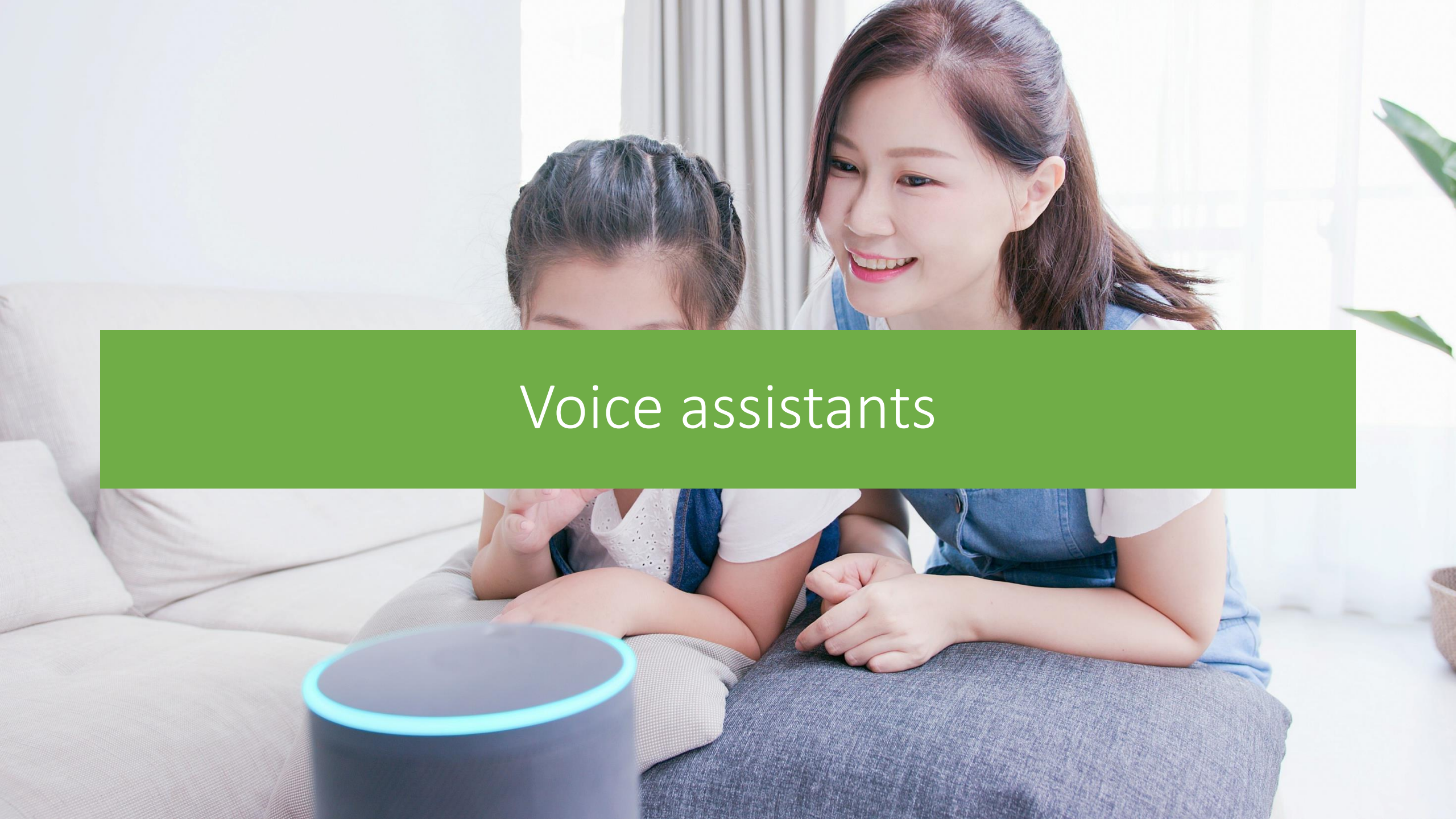

You kill kittens!
You kick puppies!



Benefits



Recommendation



Voice assistants

Robot vacuums

Navigation software

Video games

AI Benefits Top 10

Model: GPT-4

What are the top 10 benefits of modern AI?

ChatGPT

10 benefits of modern AI:

1. Improved decision-making: AI systems can analyze vast amounts of data quickly and accurately, allowing for better and faster decision-making in fields like finance, healthcare, and manufacturing.

Other Benefits

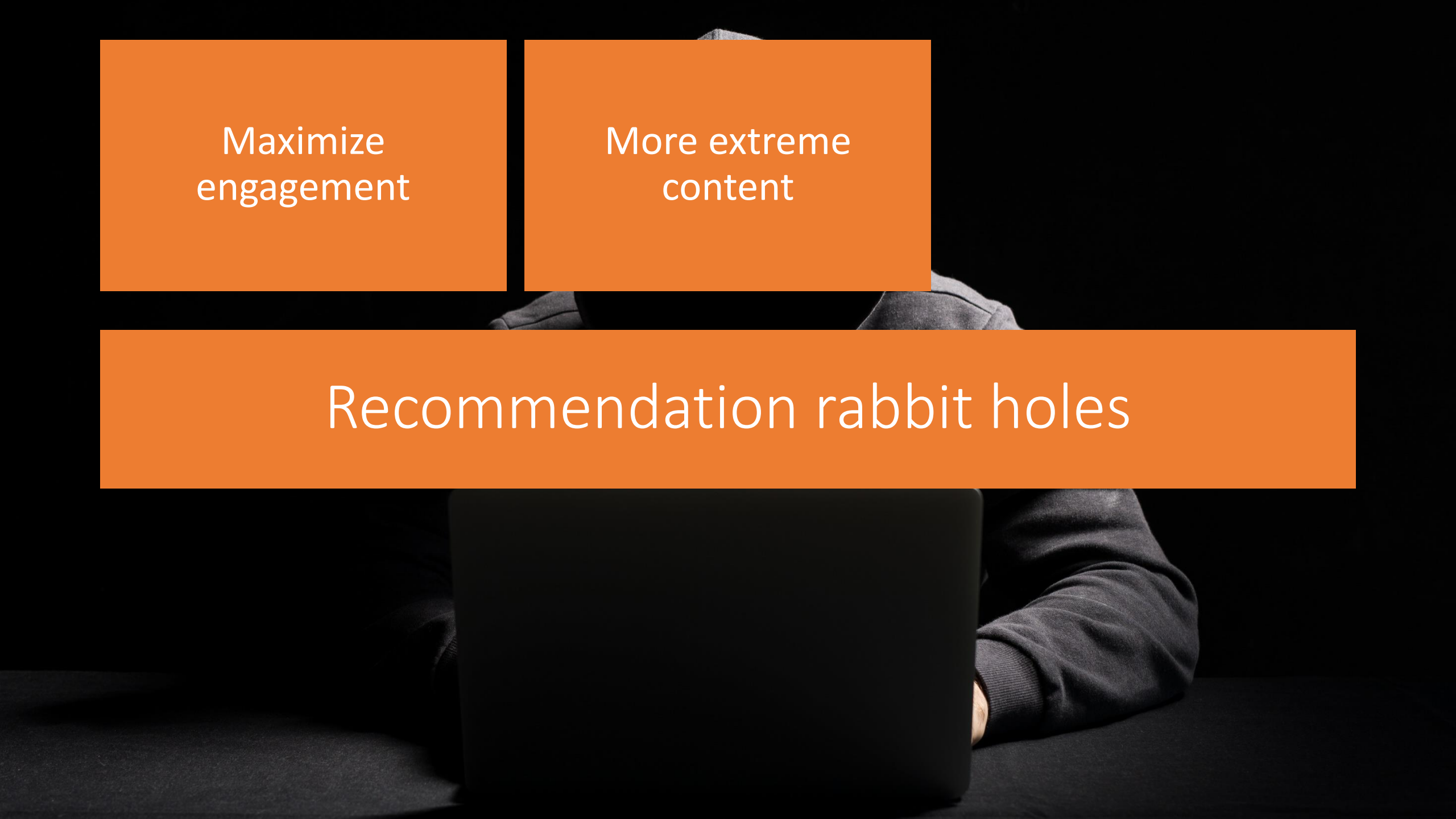
Automation	Enhanced productivity	Improved decision-making
Personalization	Cost savings	Risk mitigation
Predictive maintenance	Enhanced cybersecurity	Accessibility

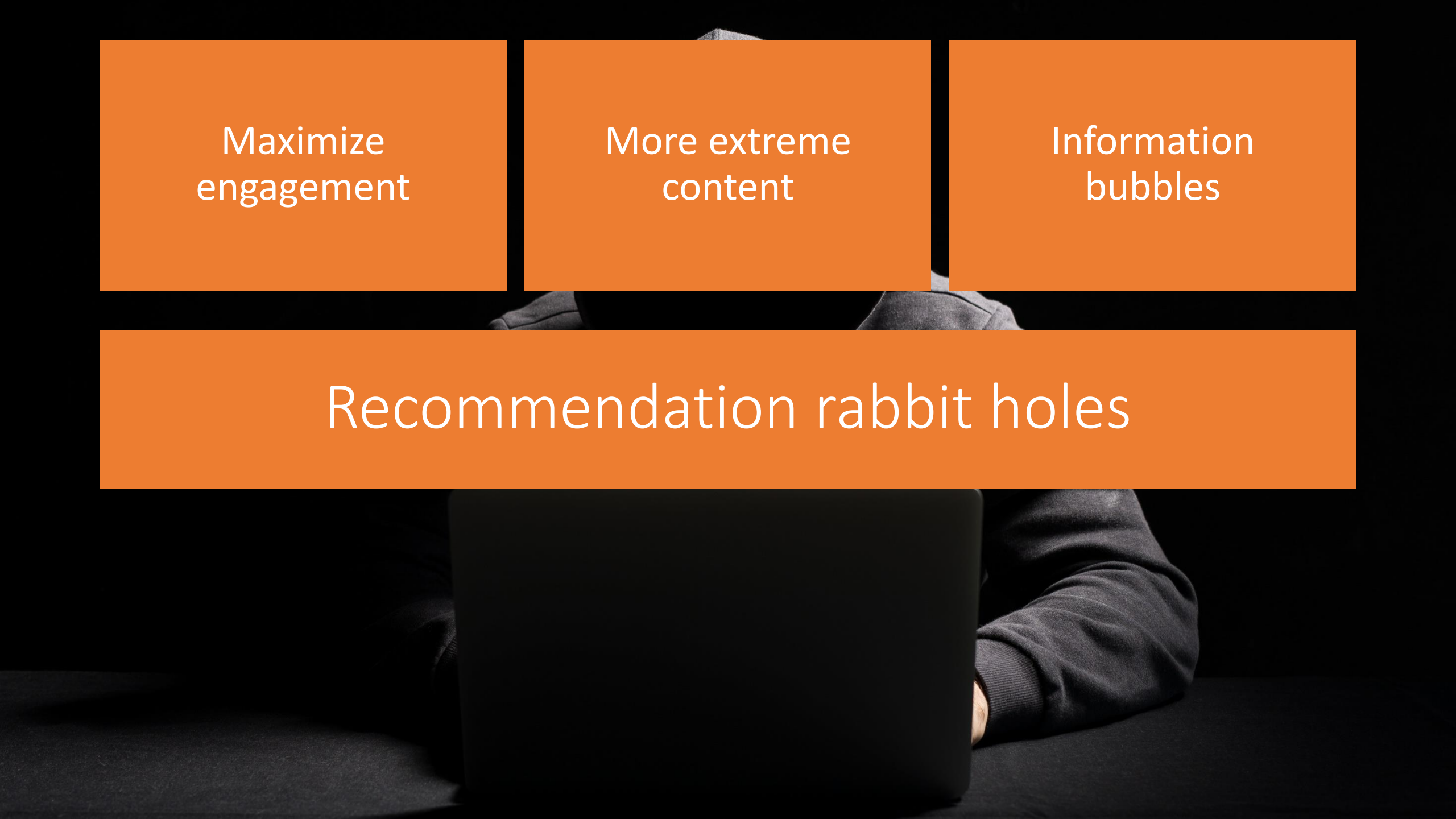
Problems

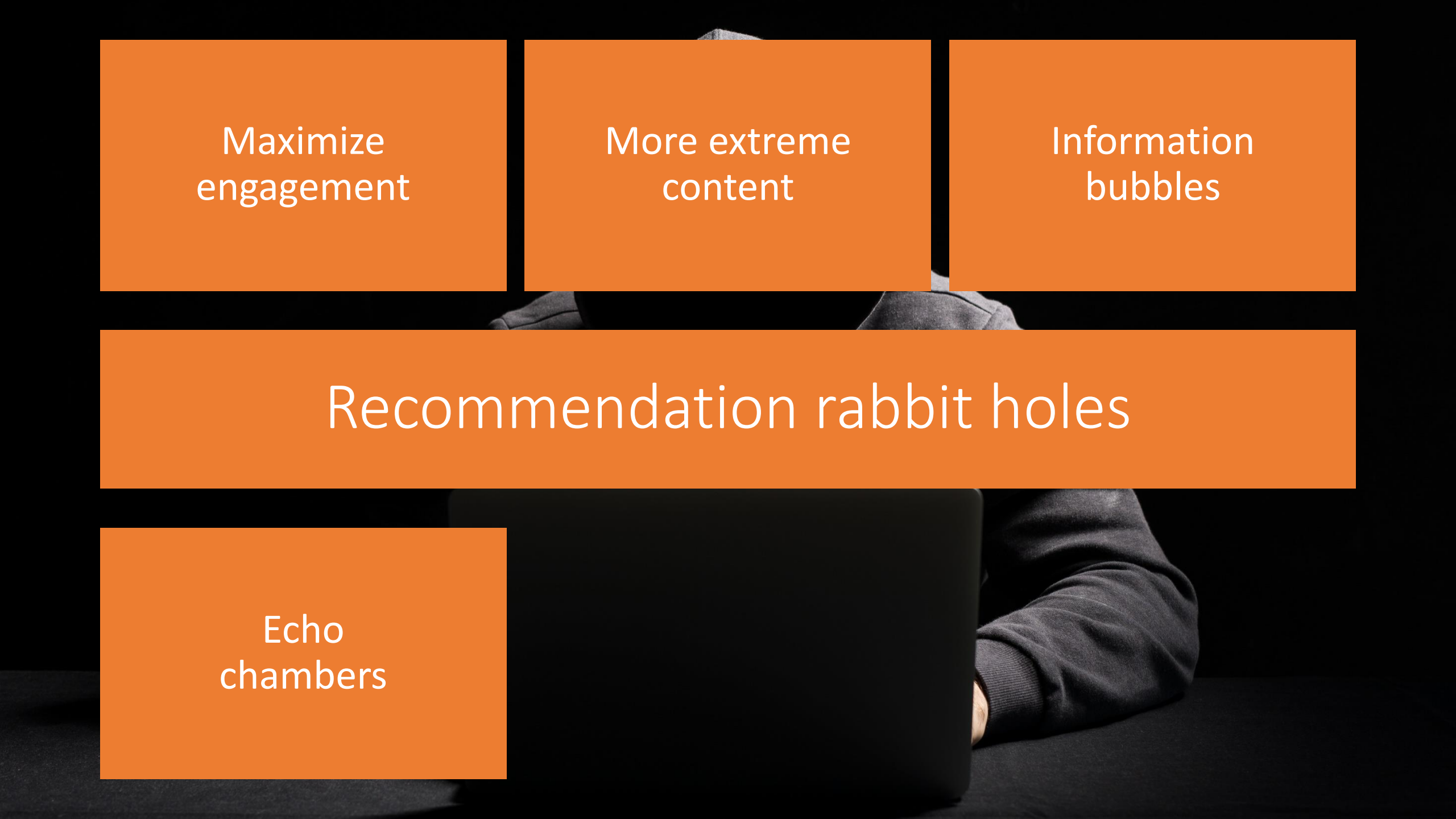
Recommendation rabbit holes

Maximize engagement

Recommendation rabbit holes

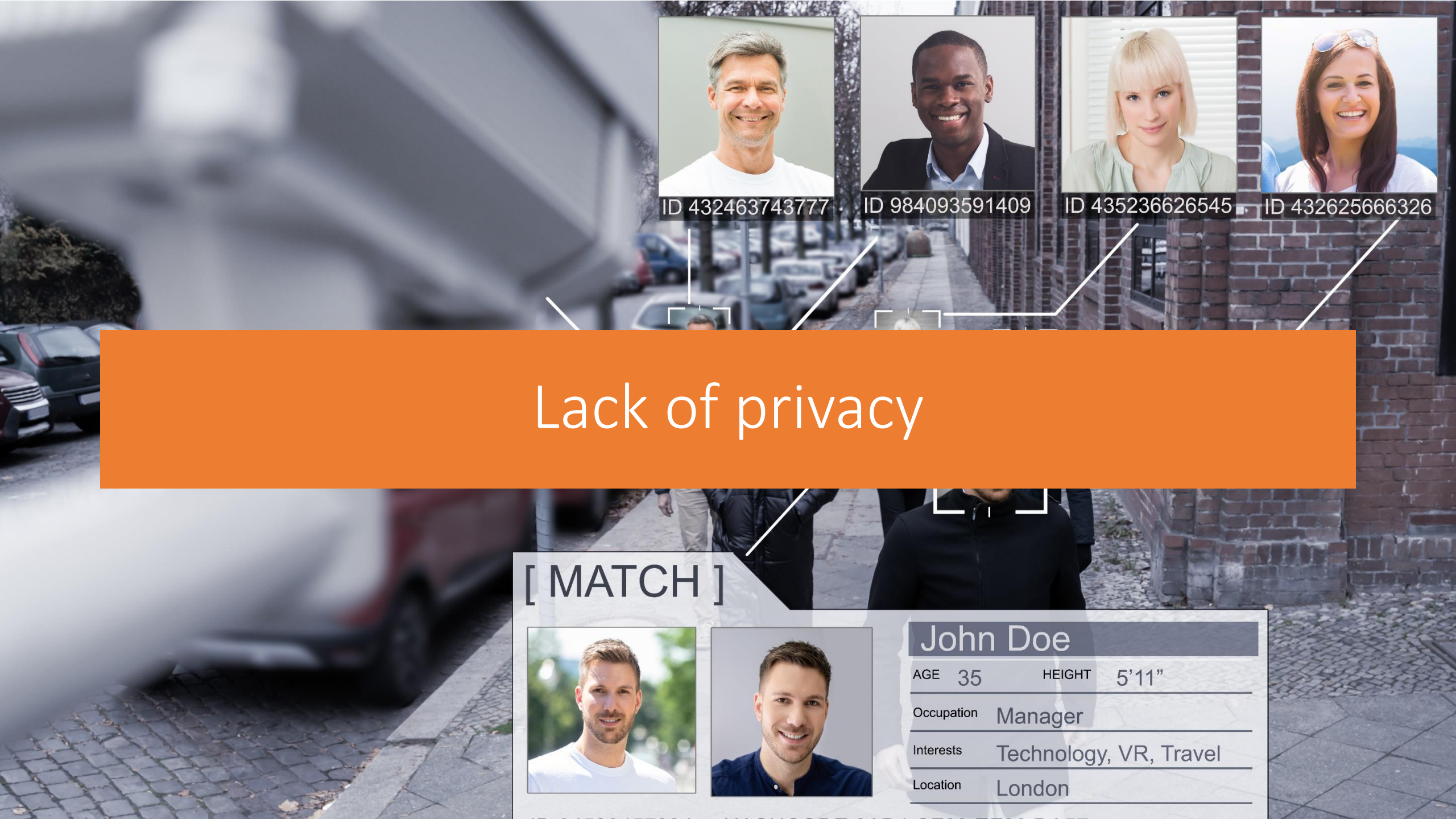

Maximize engagement
More extreme content
Recommendation rabbit holes


Maximize engagement
More extreme content
Information bubbles
Recommendation rabbit holes


Maximize engagement
More extreme content
Information bubbles
Recommendation rabbit holes
Echo chambers

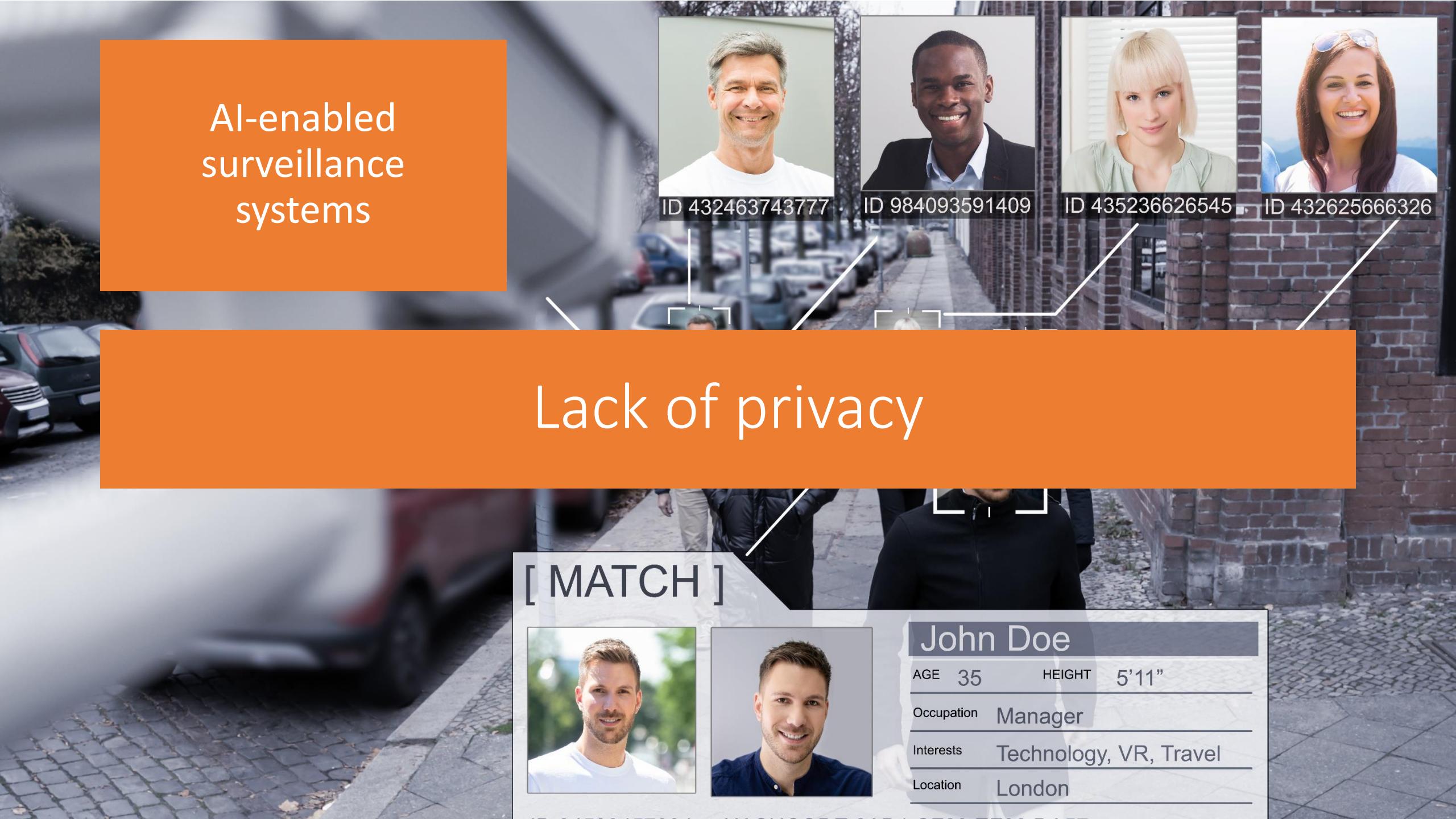
Maximize engagement
More extreme content
Information bubbles
Recommendation rabbit holes
Echo chambers
Conspiratorial thinking

Maximize engagement
More extreme content
Information bubbles
Recommendation rabbit holes
Echo chambers
Conspiratorial thinking
Internet addictions



Lack of privacy

AI-enabled surveillance systems



Lack of privacy

AI-enabled surveillance systems

Social media predicts private information

Lack of privacy

AI-enabled surveillance systems

Social media predicts private information

Shopping data predicts pregnancy

Lack of privacy

[MATCH]

John Doe

AGE	35	HEIGHT	5'11"
Occupation	Manager		
Interests	Technology, VR, Travel		
Location	London		

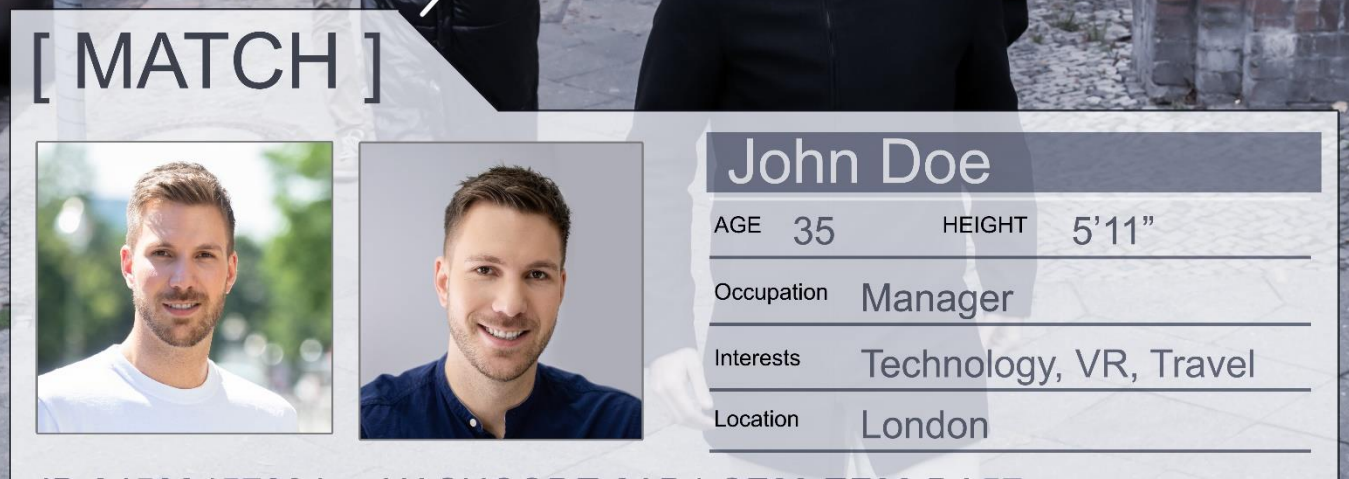
AI-enabled surveillance systems

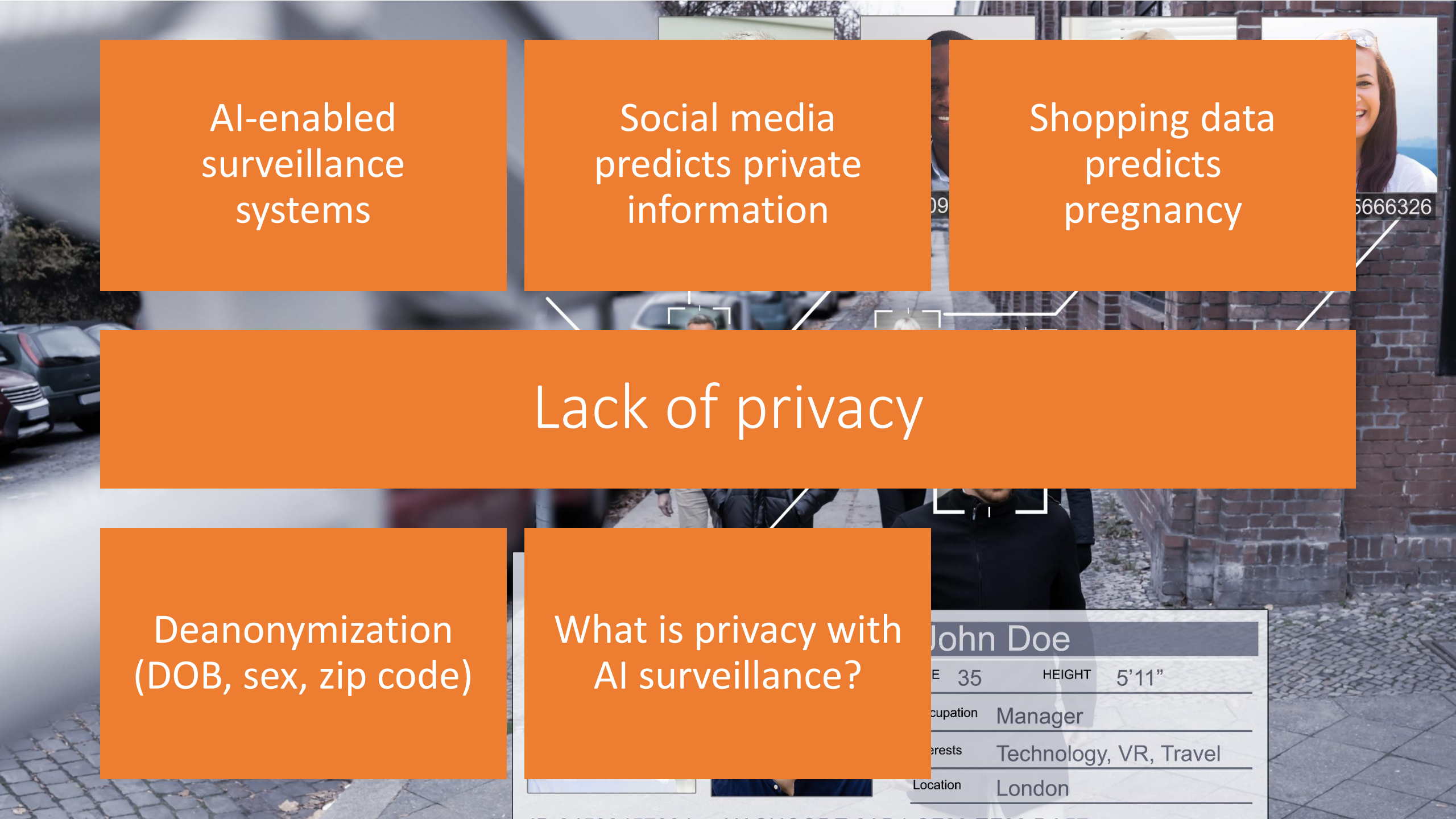
Social media predicts private information

Shopping data predicts pregnancy

Lack of privacy

Deanonymization (DOB, sex, zip code)





AI-enabled surveillance systems

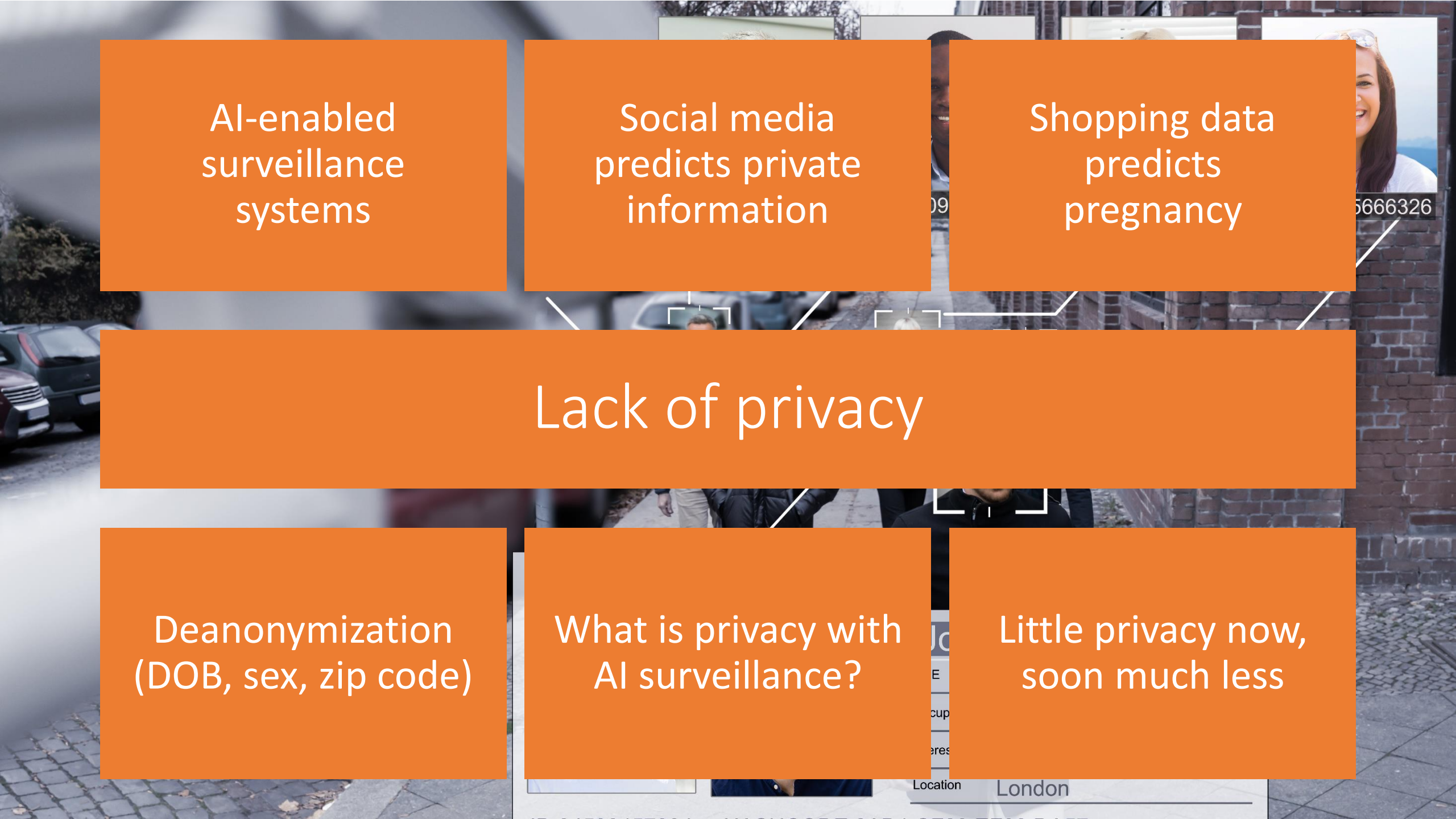
Social media predicts private information

Shopping data predicts pregnancy

Lack of privacy

Deanonymization (DOB, sex, zip code)

What is privacy with AI surveillance?



AI-enabled surveillance systems

Social media predicts private information

Shopping data predicts pregnancy

Lack of privacy

Deanonymization (DOB, sex, zip code)

What is privacy with AI surveillance?

Little privacy now, soon much less

Algorithmic bias

Racially biased criminal risk scores

Algorithmic bias

Racially biased criminal risk scores

Gender biased resume screening

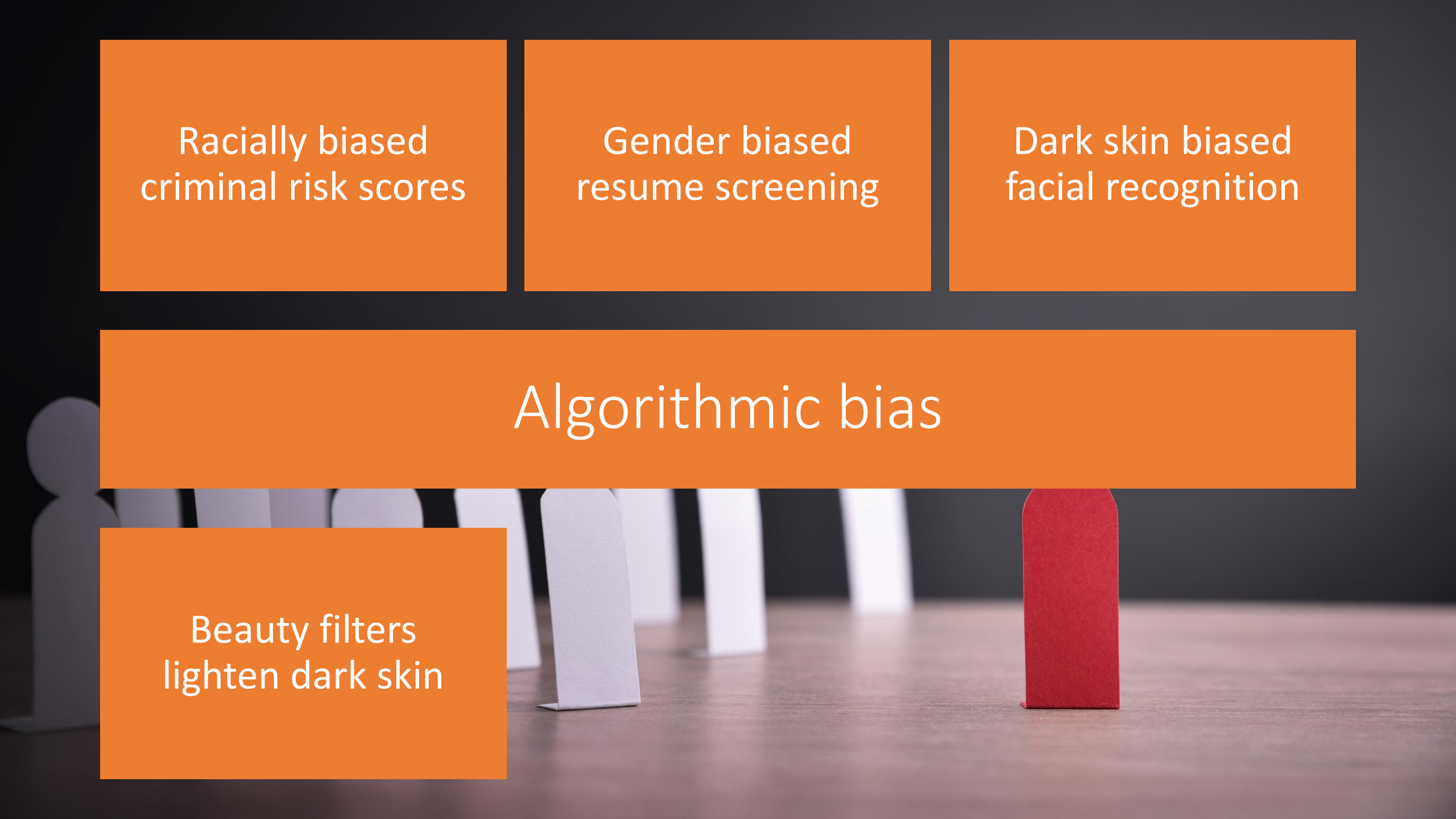
Algorithmic bias

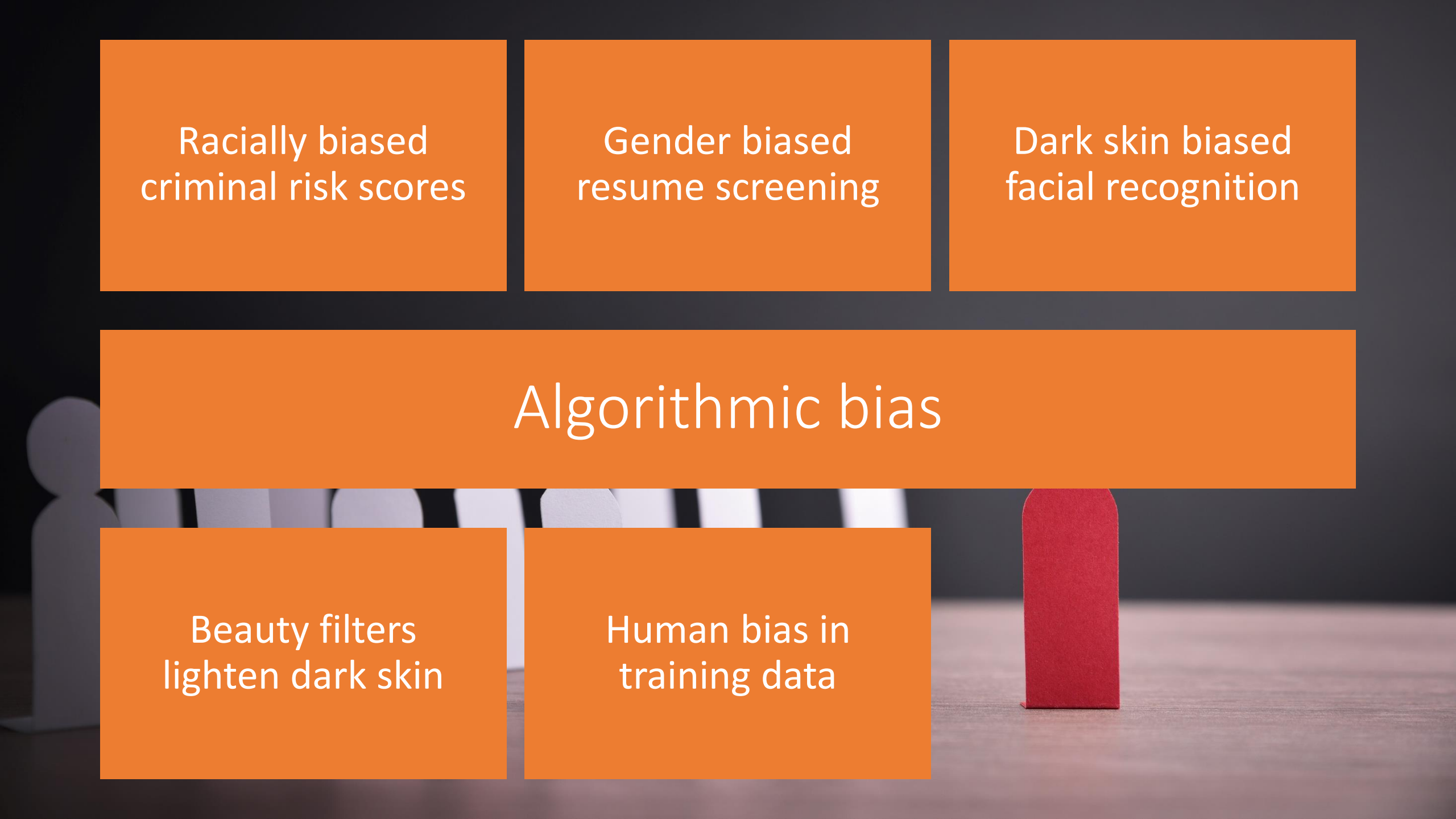
Racially biased criminal risk scores

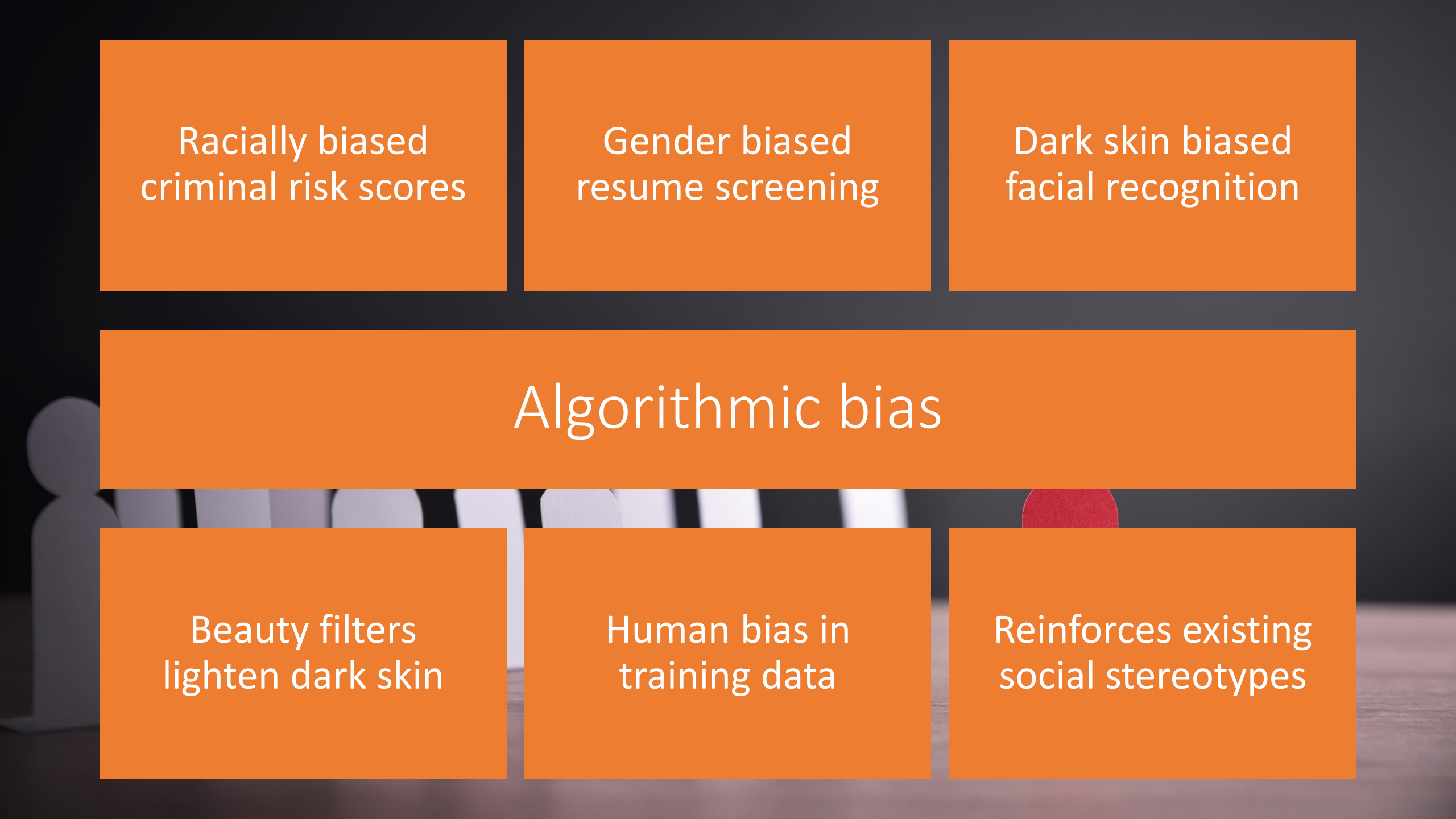
Gender biased resume screening

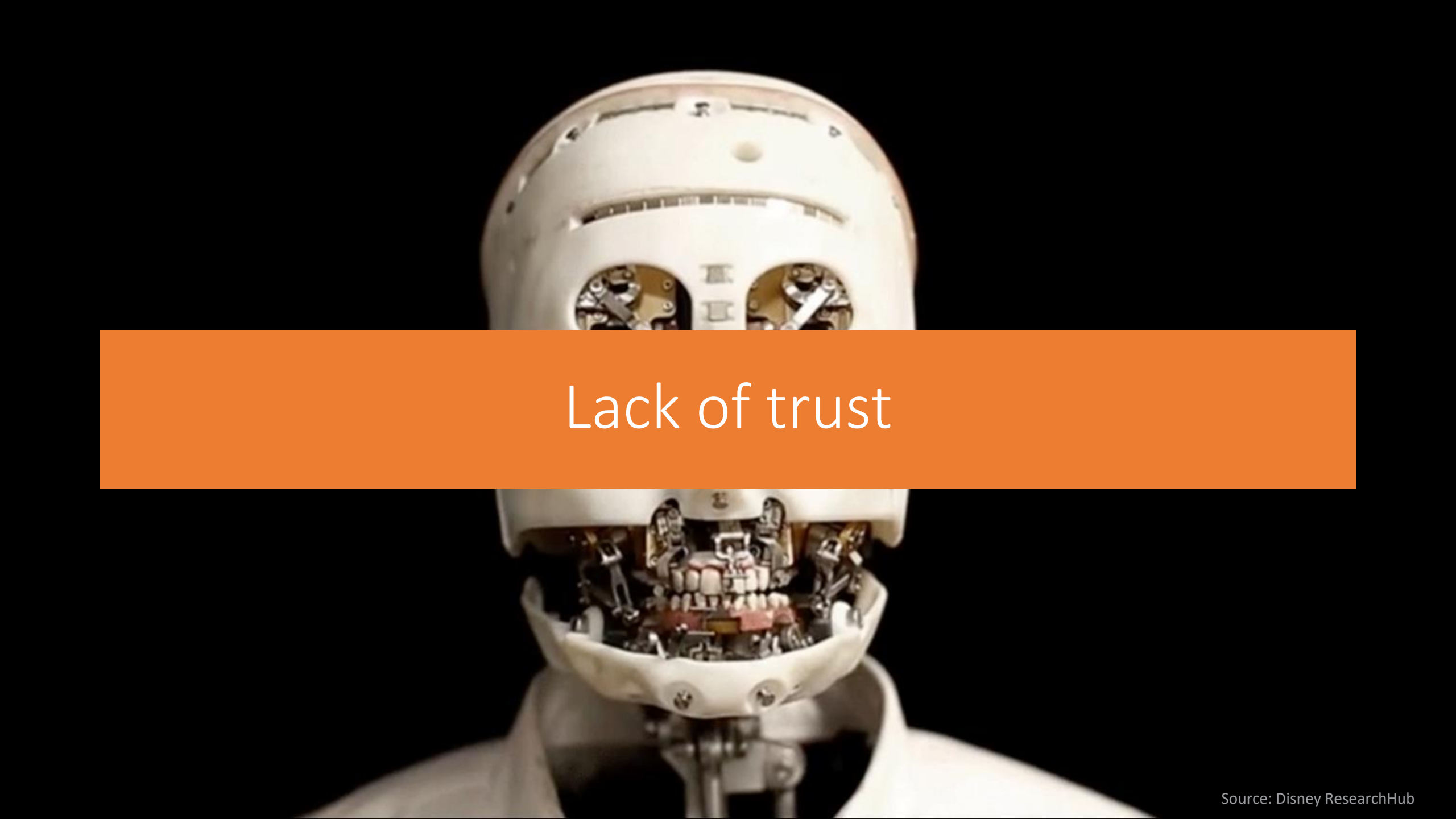
Dark skin biased facial recognition

Algorithmic bias


Racially biased criminal risk scores
Gender biased resume screening
Dark skin biased facial recognition
Algorithmic bias
Beauty filters lighten dark skin


Racially biased criminal risk scores
Gender biased resume screening
Dark skin biased facial recognition
Algorithmic bias
Beauty filters lighten dark skin
Human bias in training data

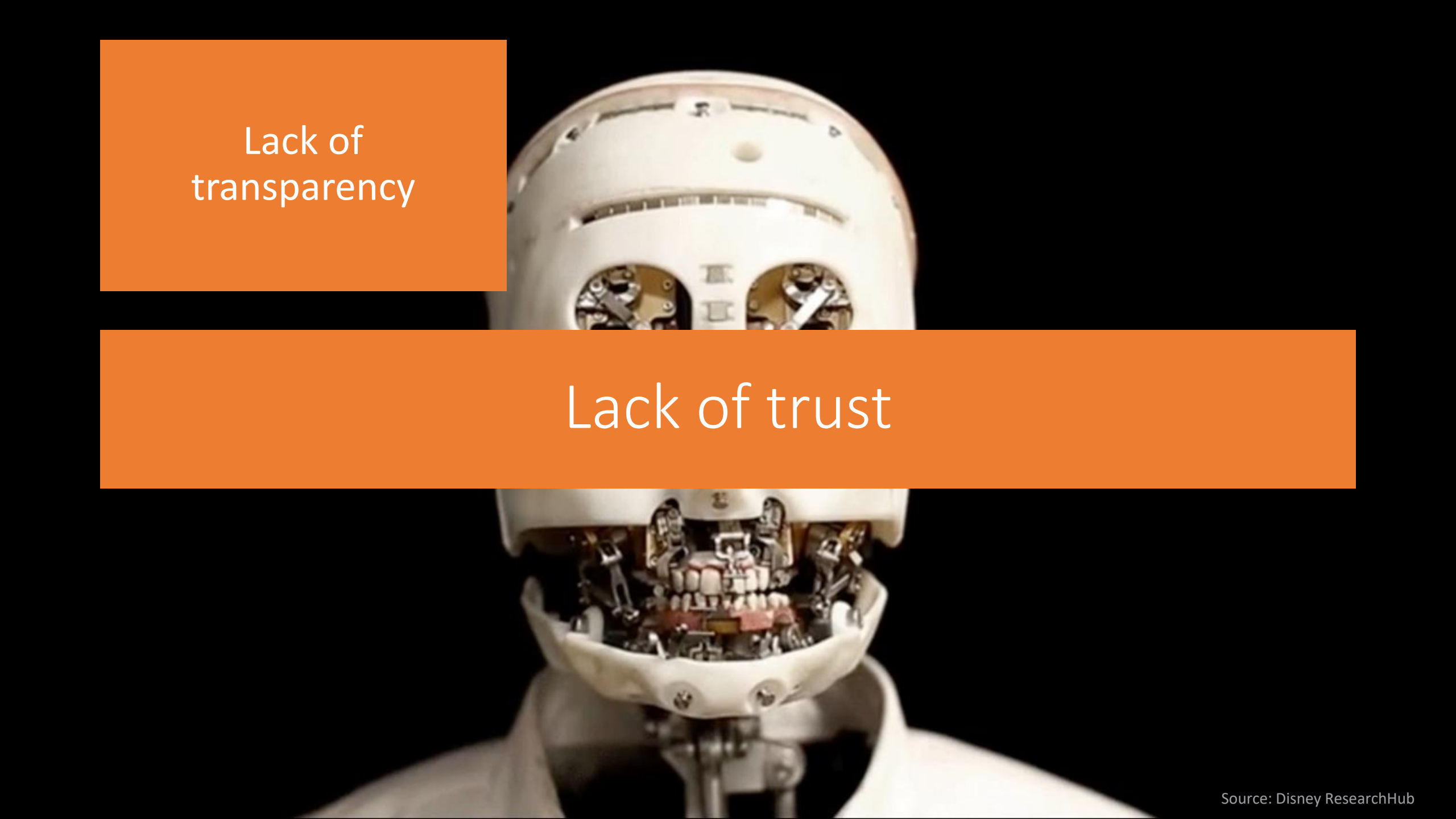

Racially biased criminal risk scores
Gender biased resume screening
Dark skin biased facial recognition
Algorithmic bias
Beauty filters lighten dark skin
Human bias in training data
Reinforces existing social stereotypes



Lack of trust

Source: Disney ResearchHub

Lack of transparency

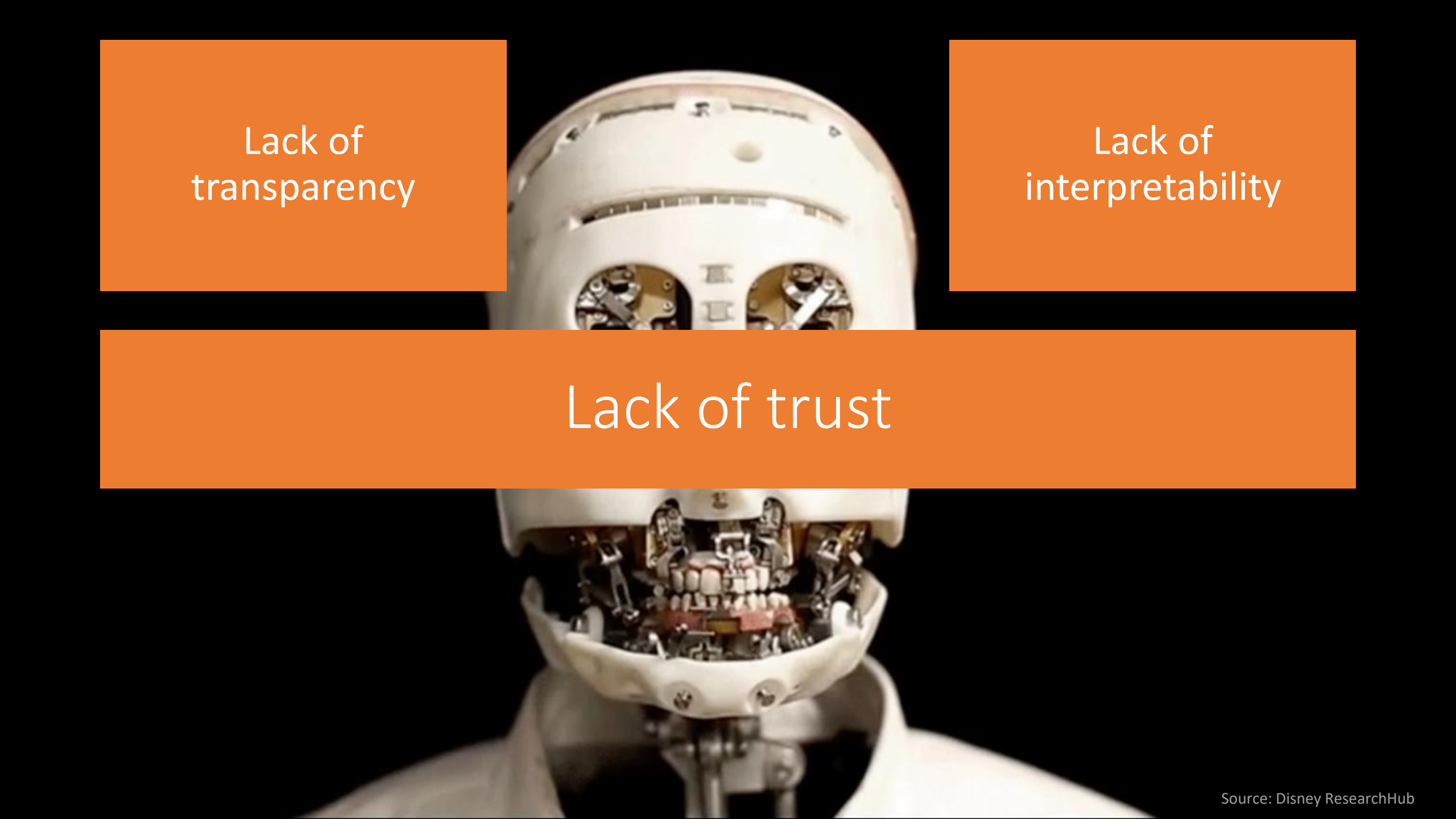


Lack of trust

Source: Disney ResearchHub

Lack of transparency

Lack of interpretability

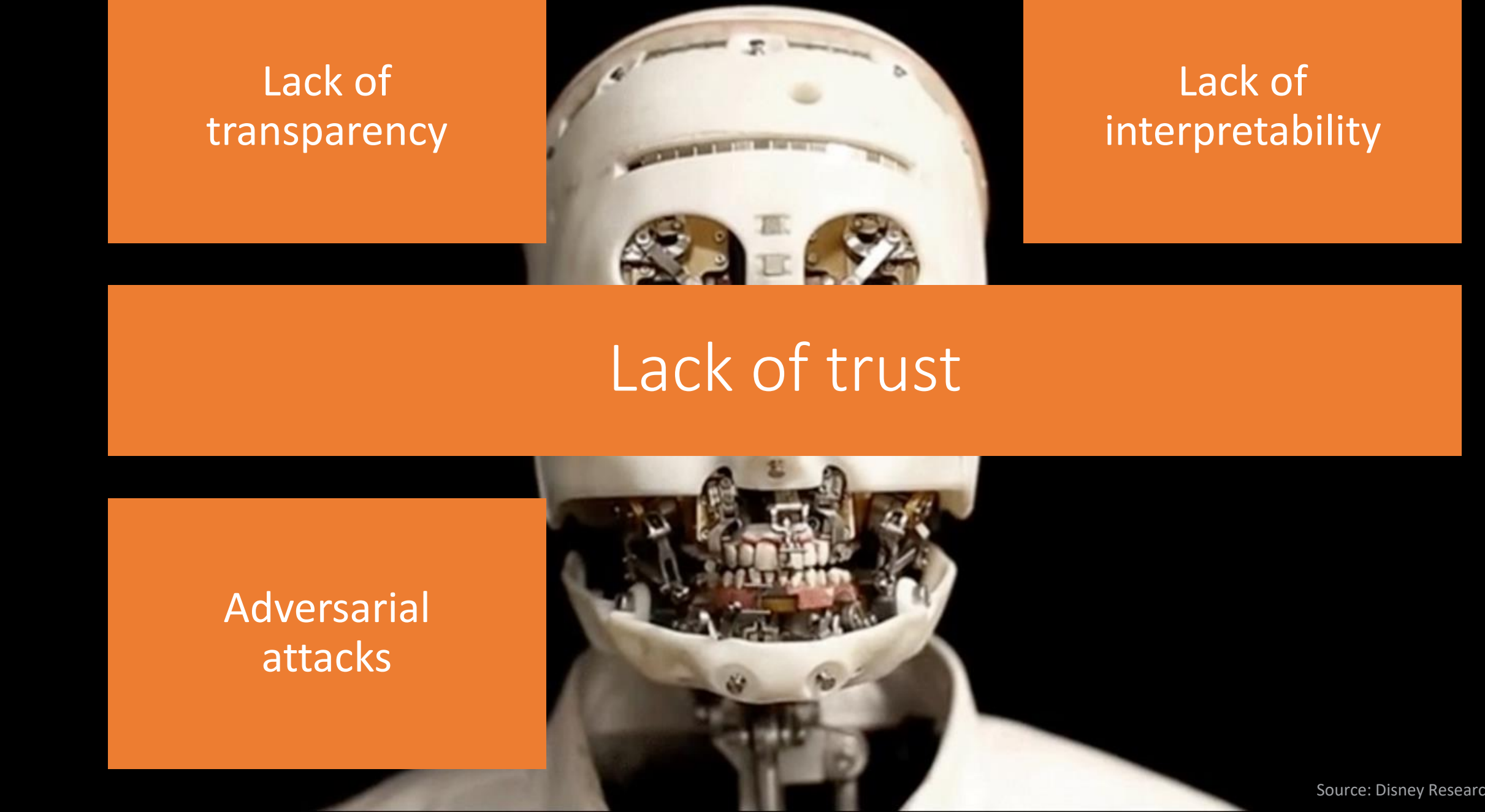


Lack of trust

Source: Disney ResearchHub

Lack of transparency

Lack of interpretability



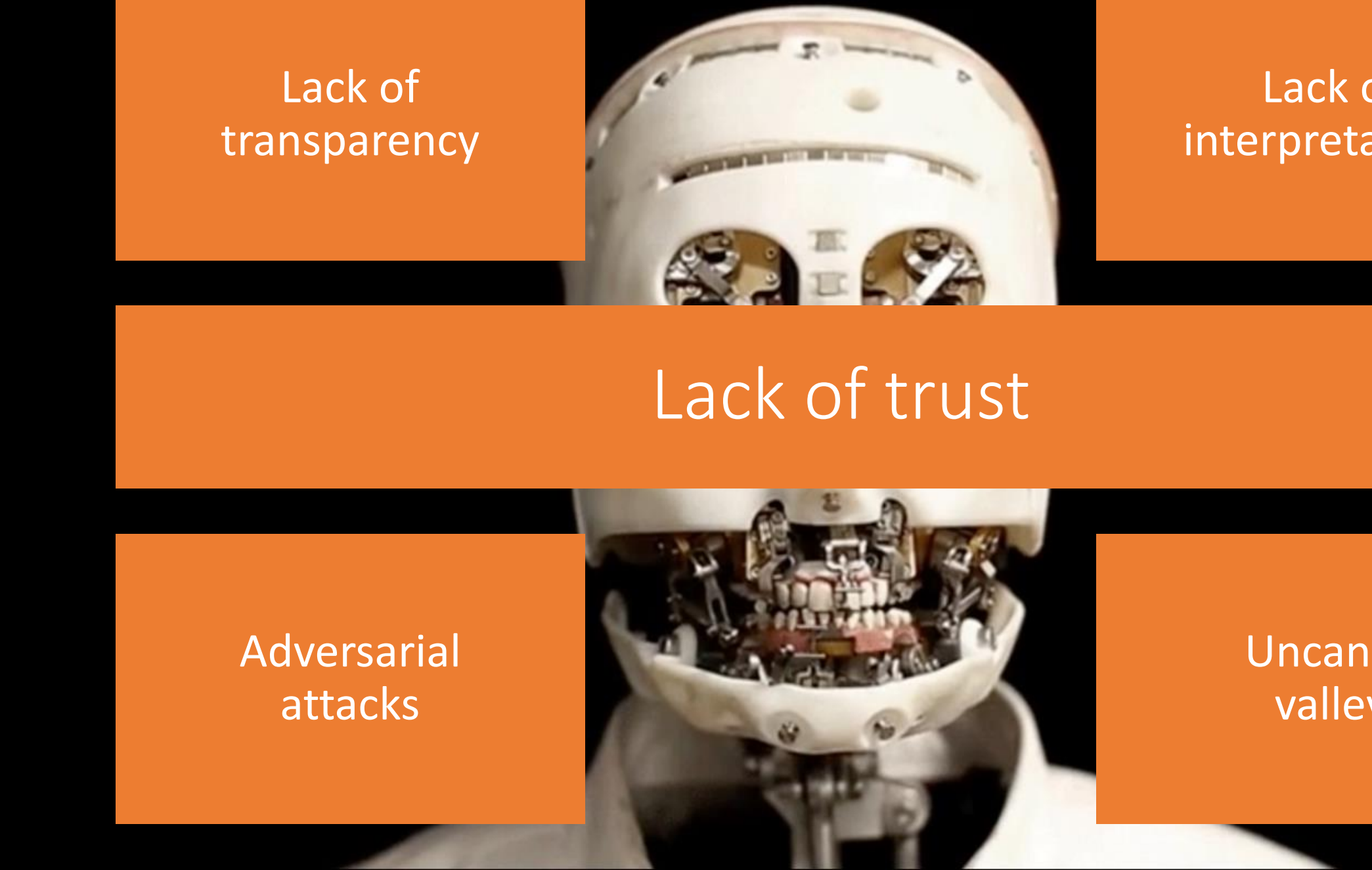
Lack of trust

Adversarial attacks

Source: Disney ResearchHub

Lack of transparency

Lack of interpretability



Lack of trust

Adversarial attacks

Uncanny valley

Source: Disney ResearchHub

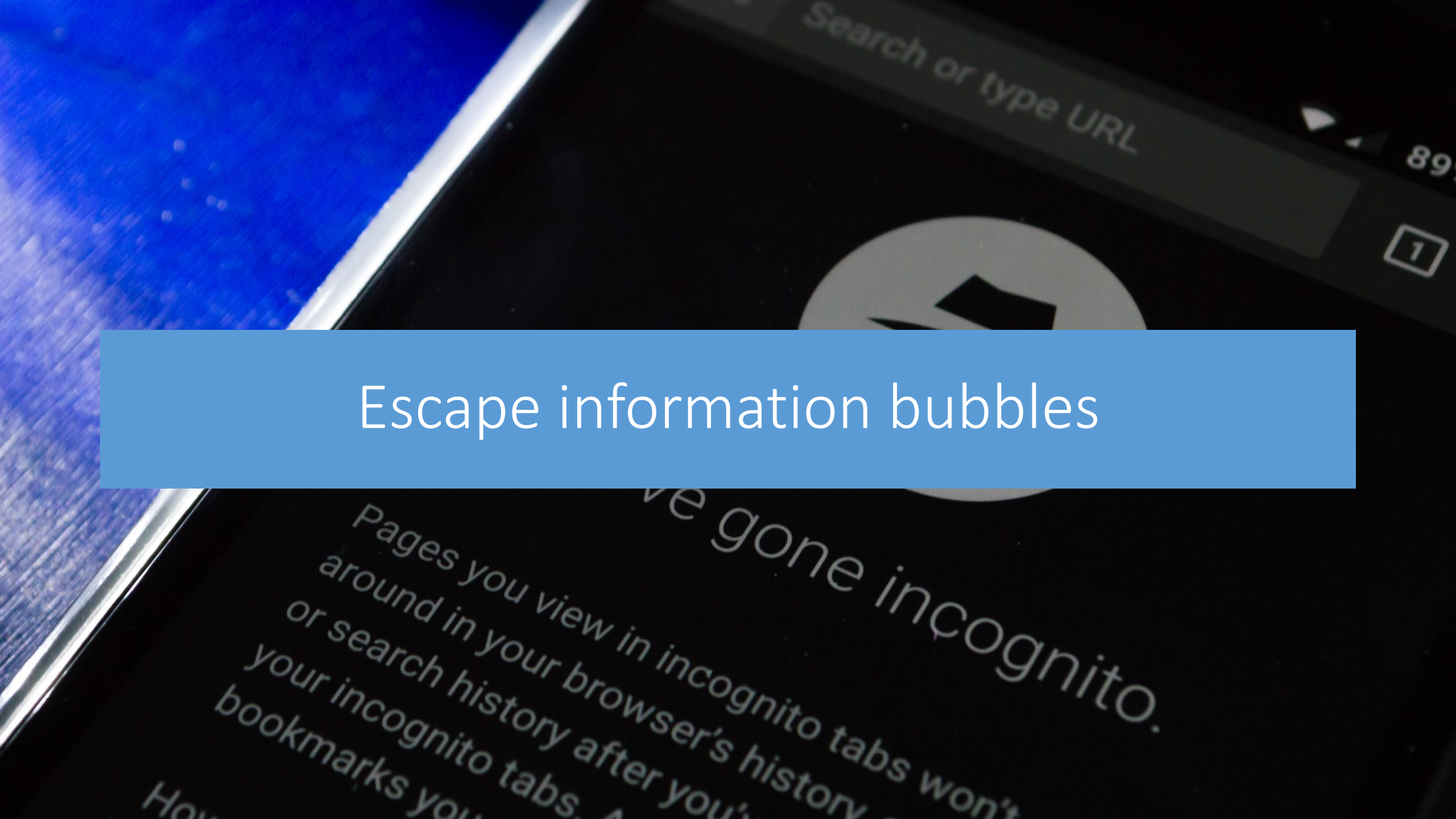
Hallucination

Source: Pouff - Grocery Trip

Misinformation

Lack of AI literacy

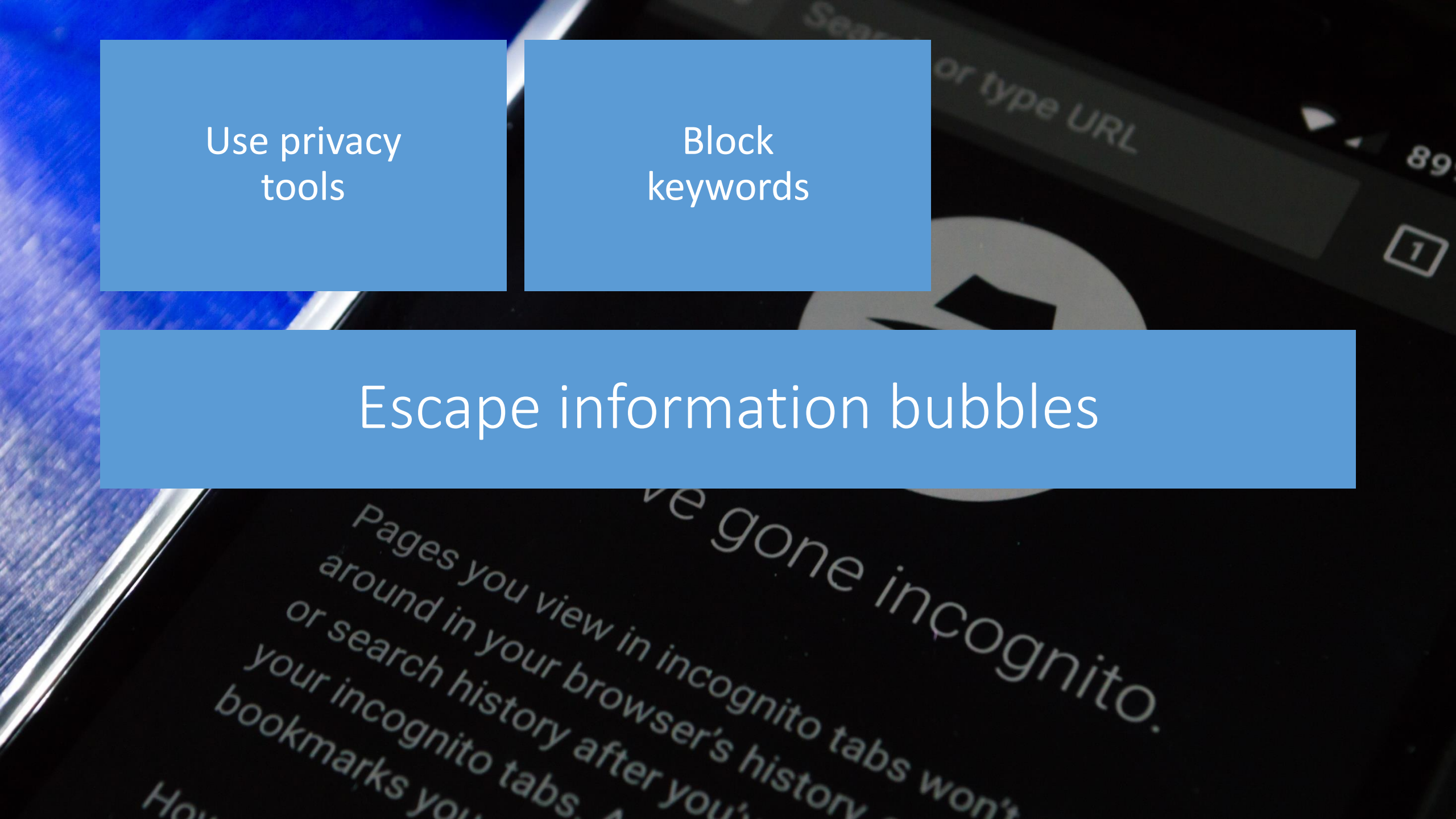
Solutions



Escape information bubbles

Use privacy tools

Escape information bubbles

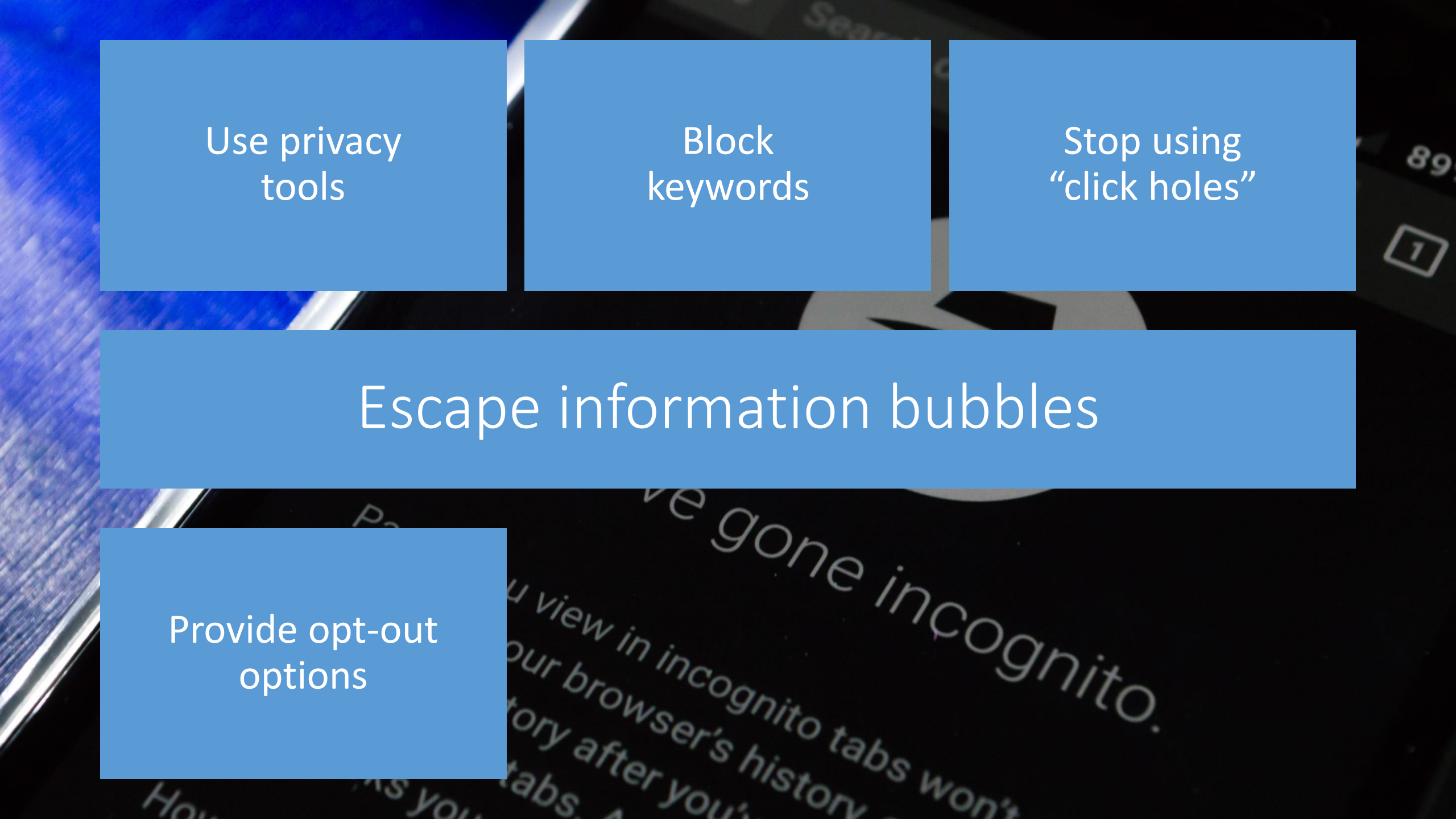


Use privacy tools

Block keywords

Escape information bubbles

Use privacy tools
Block keywords
Stop using “click holes”
Escape information bubbles
ve gone incognito.
Pages you view in incognito tabs won't
around in your browser's history
or search history after you
your incognito tabs
bookmarks



Use privacy tools

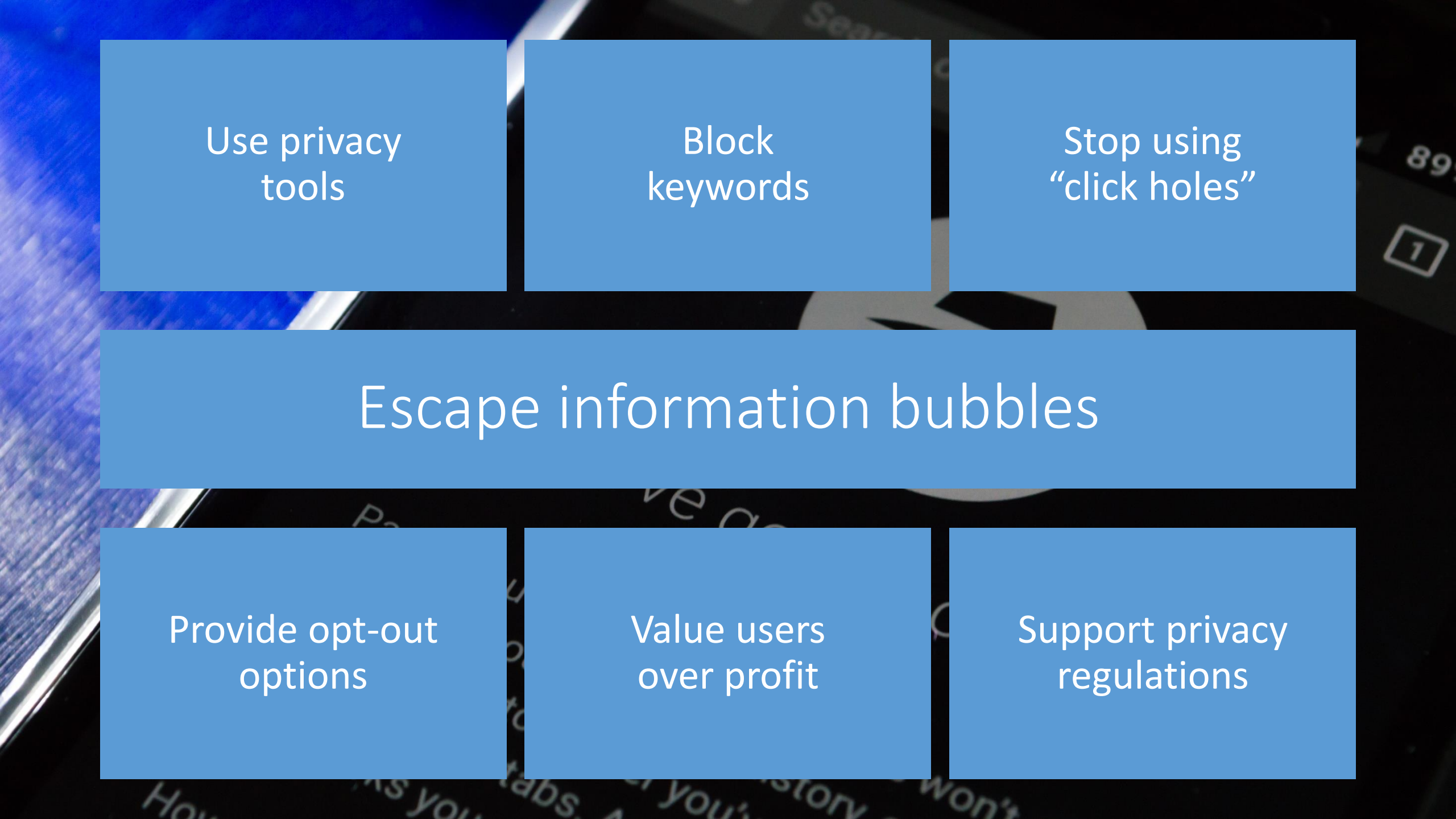
Block keywords

Stop using “click holes”

Escape information bubbles

Provide opt-out options

Use privacy tools
Block keywords
Stop using “click holes”
Escape information bubbles
Provide opt-out options
Value users over profit



Use privacy tools

Block keywords

Stop using “click holes”

Escape information bubbles

Provide opt-out options

Value users over profit

Support privacy regulations

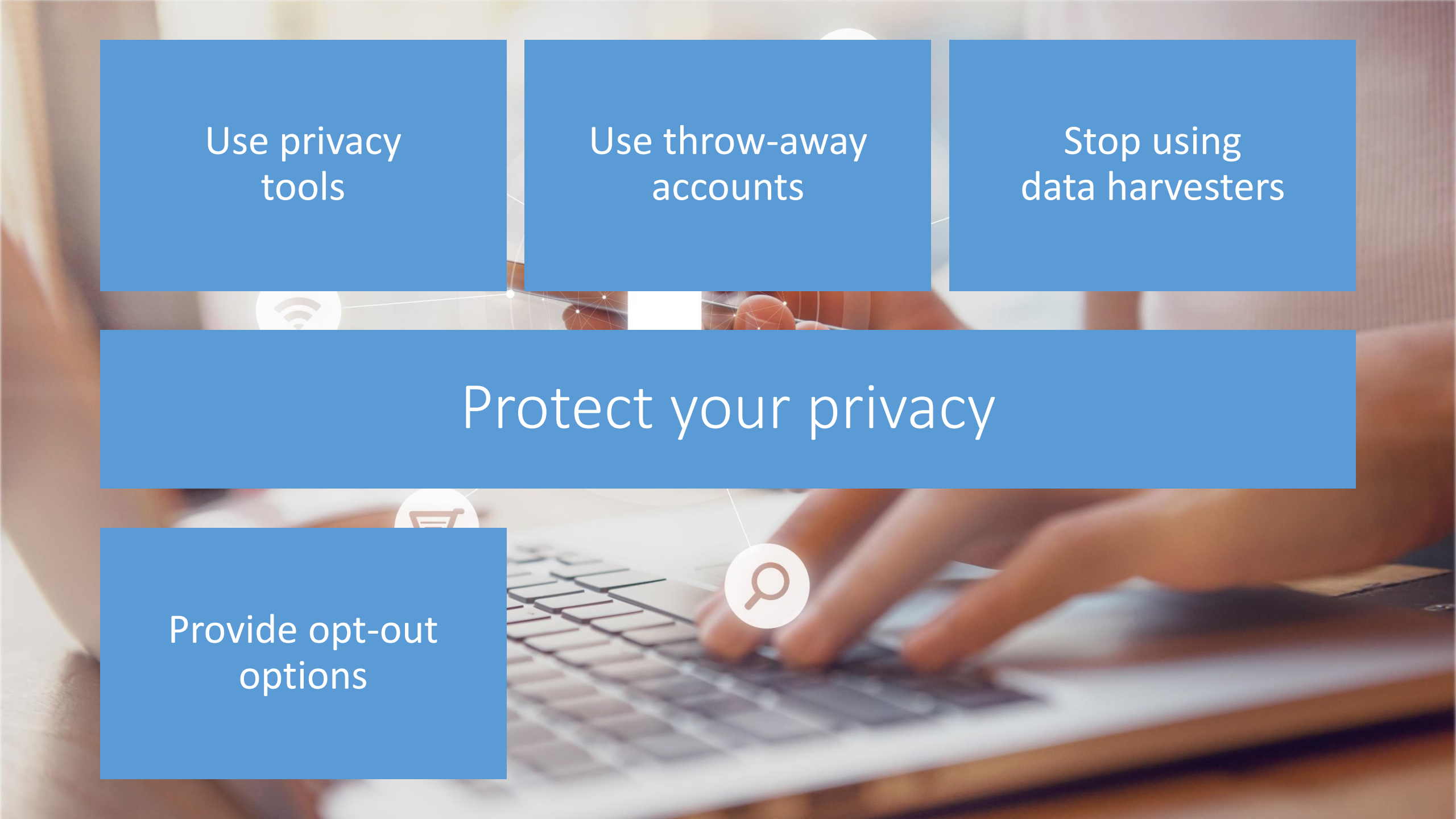
Protect your privacy

Use privacy tools

Protect your privacy

Use privacy tools
Use throw-away accounts
Protect your privacy

Use privacy tools
Use throw-away accounts
Stop using data harvesters
Protect your privacy


Use privacy tools
Use throw-away accounts
Stop using data harvesters
Protect your privacy
Provide opt-out options

Use privacy tools
Use throw-away accounts
Stop using data harvesters
Protect your privacy
Provide opt-out options
Use strong anonymization

Use privacy tools
Use throw-away accounts
Stop using data harvesters
Protect your privacy
Provide opt-out options
Use strong anonymization
Support GDPR and CCPA

Use privacy tools
Use throw-away accounts
Stop using data harvesters
Protect your privacy
Provide opt-out options
Use strong anonymization
Support GDPR and CCPA

Eliminate bias

Use diverse datasets
Eliminate bias

Use diverse datasets
Exclude protected attributes
Eliminate bias

Use diverse datasets
Exclude protected attributes
Use bias-detection tools
Eliminate bias

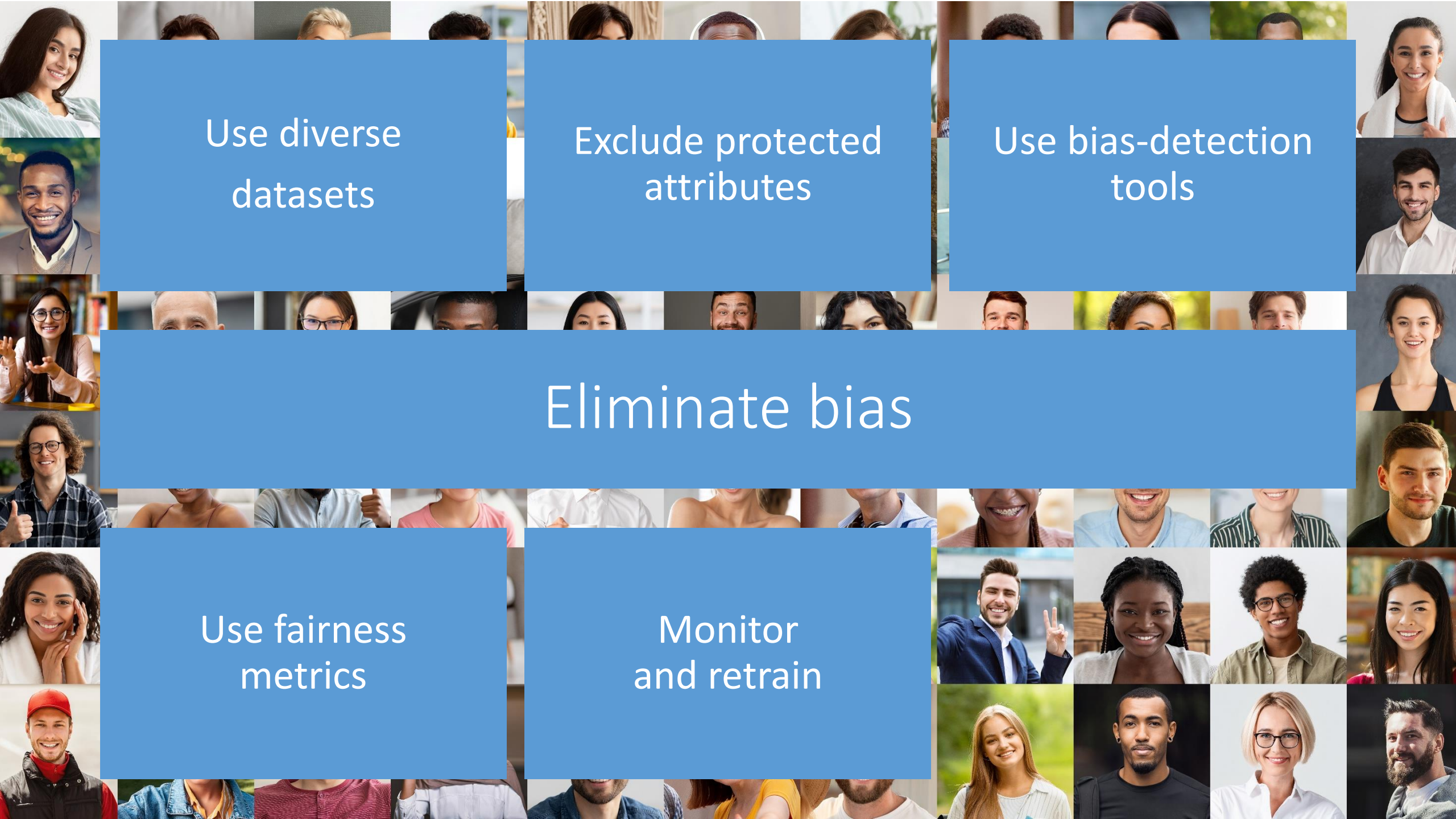
Use diverse datasets

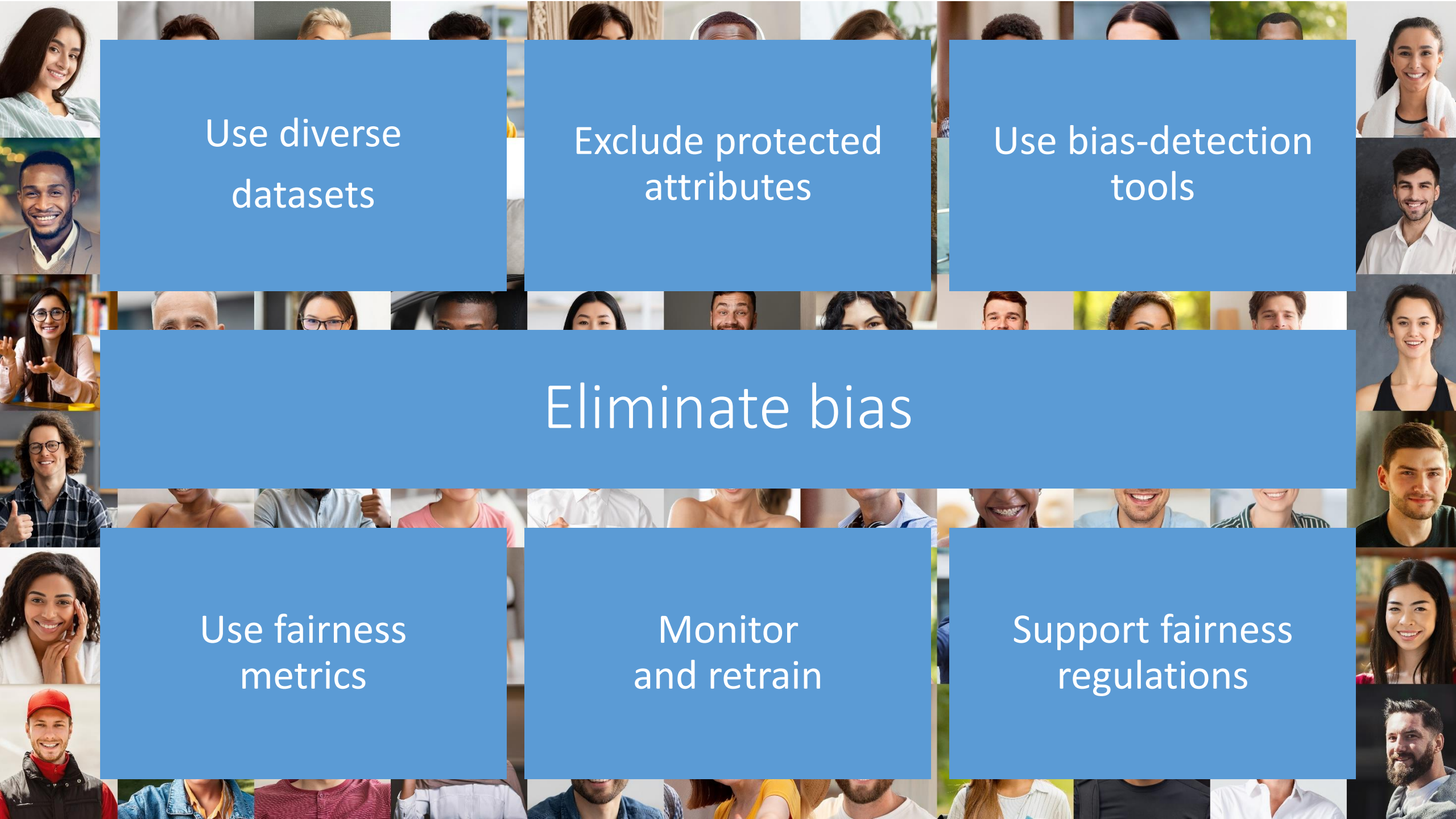
Exclude protected attributes

Use bias-detection tools

Eliminate bias

Use fairness metrics


Use diverse datasets
Exclude protected attributes
Use bias-detection tools
Eliminate bias
Use fairness metrics
Monitor and retrain


Use diverse datasets
Exclude protected attributes
Use bias-detection tools
Eliminate bias
Use fairness metrics
Monitor and retrain
Support fairness regulations

Create trustworthy AI

Choose transparency

Create trustworthy AI

Choose transparency

Keep it simple

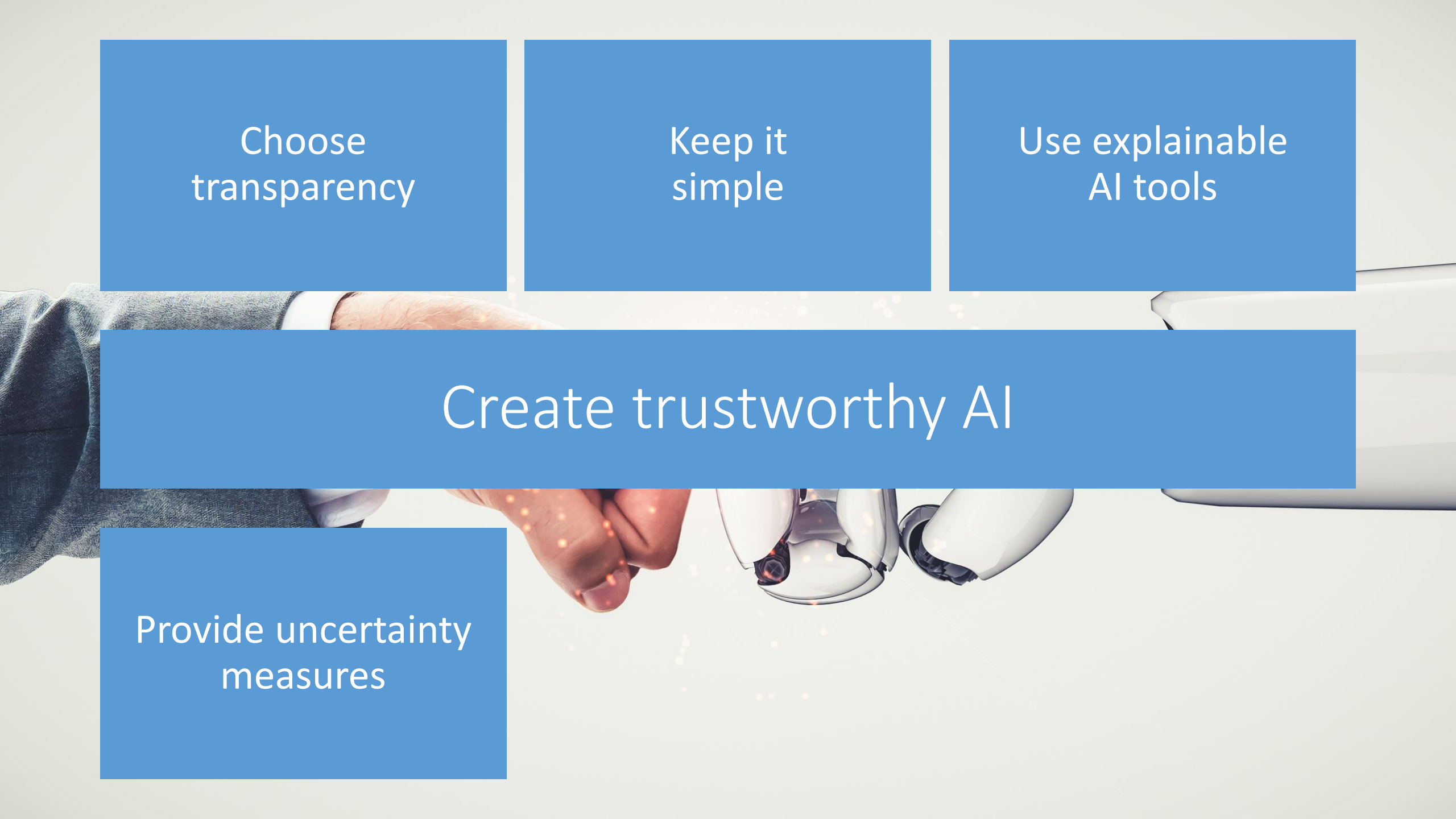
Create trustworthy AI

Choose transparency

Keep it simple

Use explainable AI tools

Create trustworthy AI



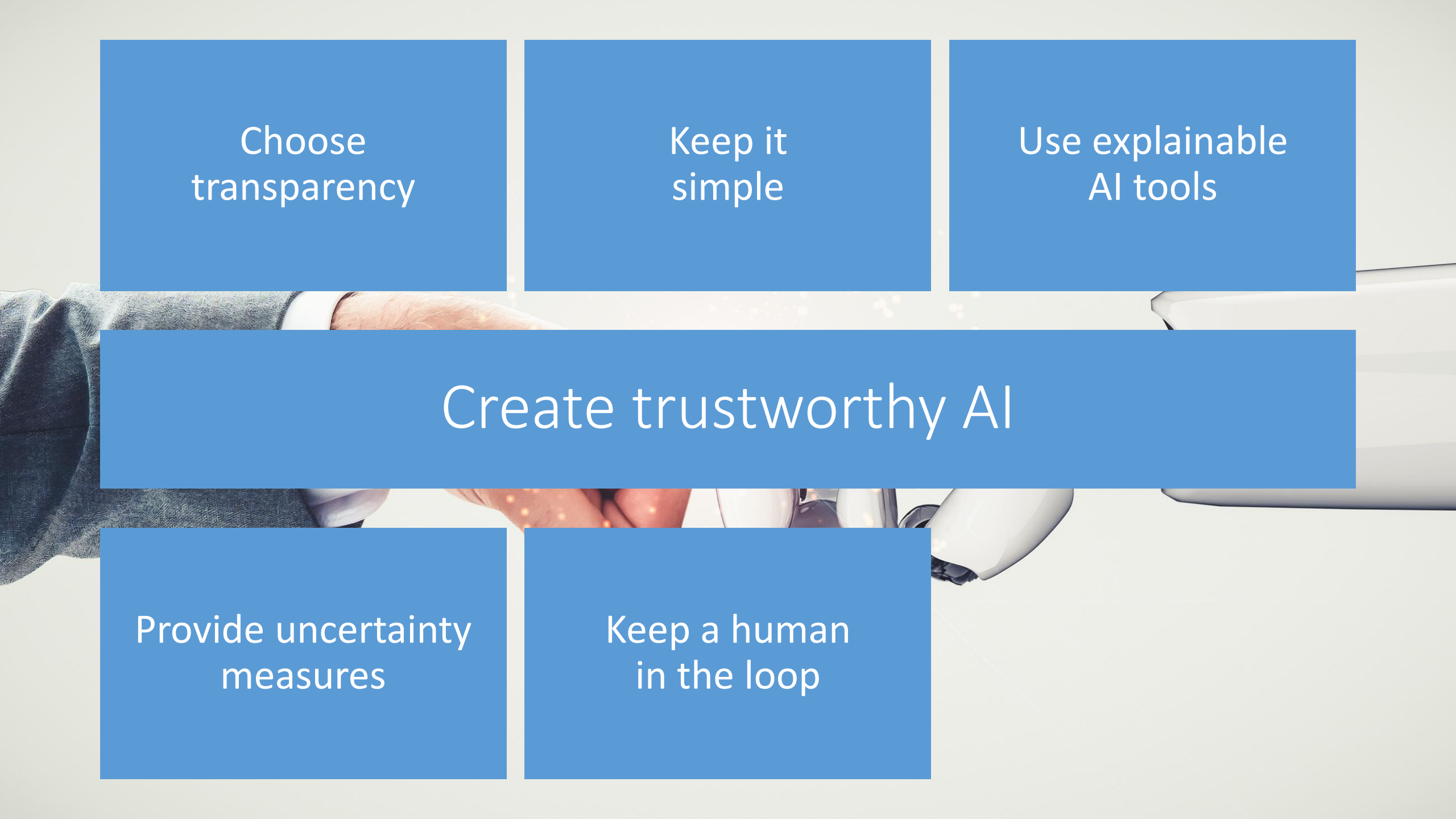
Choose transparency

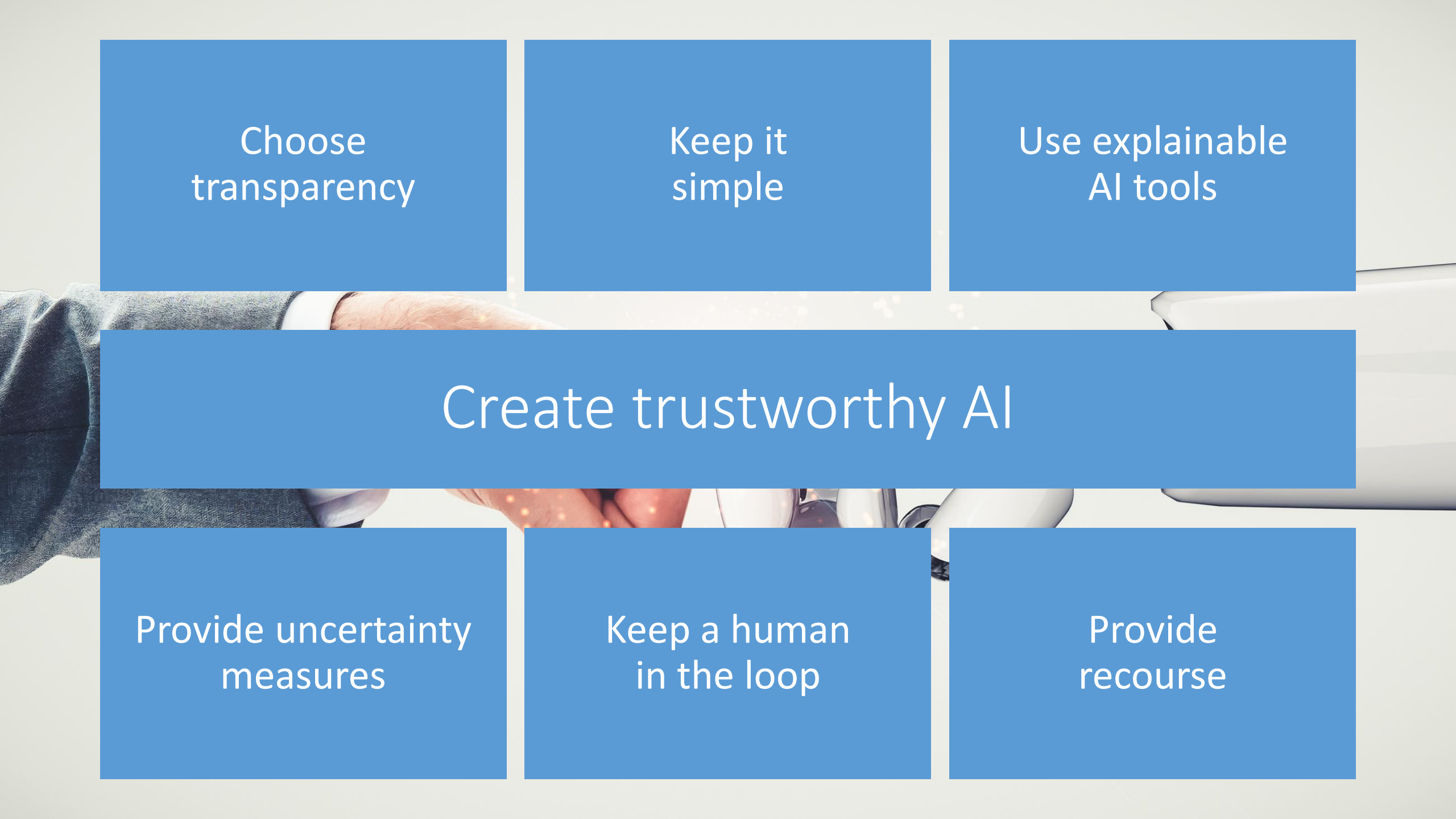
Keep it simple

Use explainable AI tools

Create trustworthy AI

Provide uncertainty measures


Choose transparency
Keep it simple
Use explainable AI tools
Create trustworthy AI
Provide uncertainty measures
Keep a human in the loop


Choose transparency
Keep it simple
Use explainable AI tools
Create trustworthy AI
Provide uncertainty measures
Keep a human in the loop
Provide recourse

Fight misinformation

Who is the author?
Fight misinformation

Who is the author?
Who is the publisher?
Fight misinformation

Who is the author?
Who is the publisher?
What are their claims?
Fight misinformation

Who is the author?
Who is the publisher?
What are their claims?
Fight misinformation
How strong is their evidence?

Who is the author?
Who is the publisher?
What are their claims?
Fight misinformation
How strong is their evidence?
What are their sources?

Who is the author?
Who is the publisher?
What are their claims?
Fight misinformation
How strong is their evidence?
What are their sources?
Be skeptical but open-minded

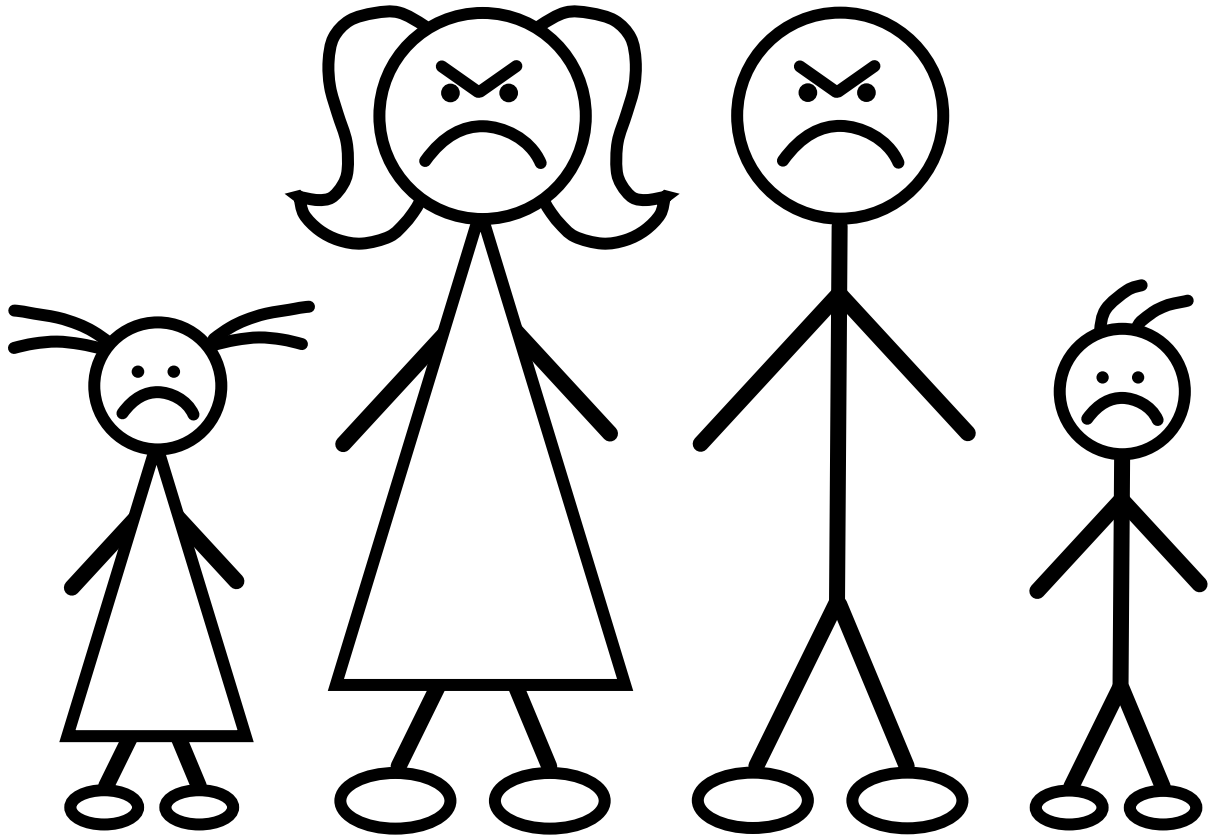
Get off social media

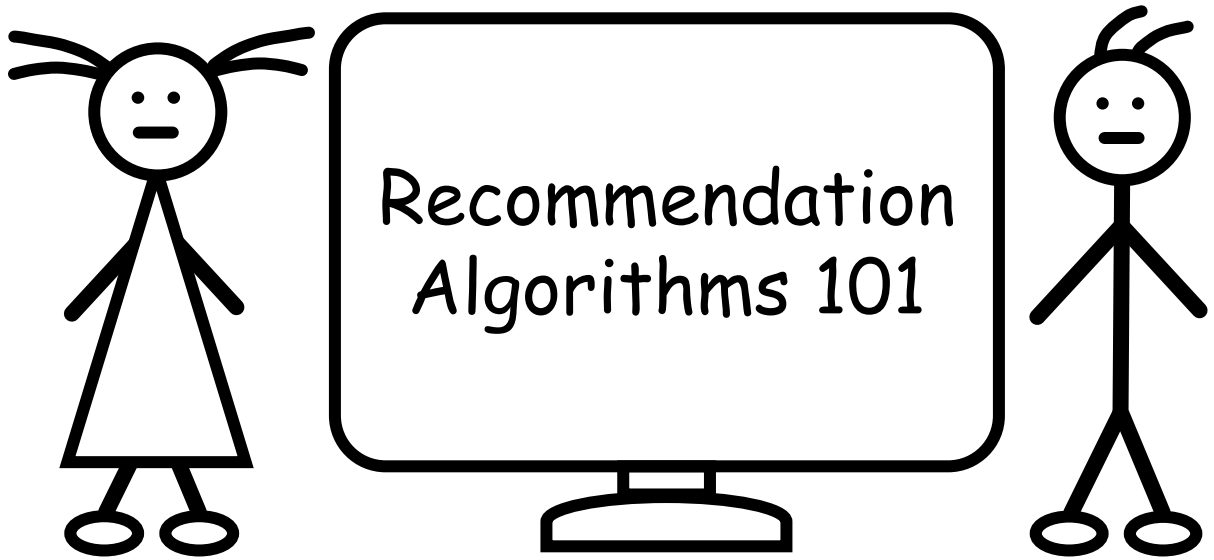
Talk to other people

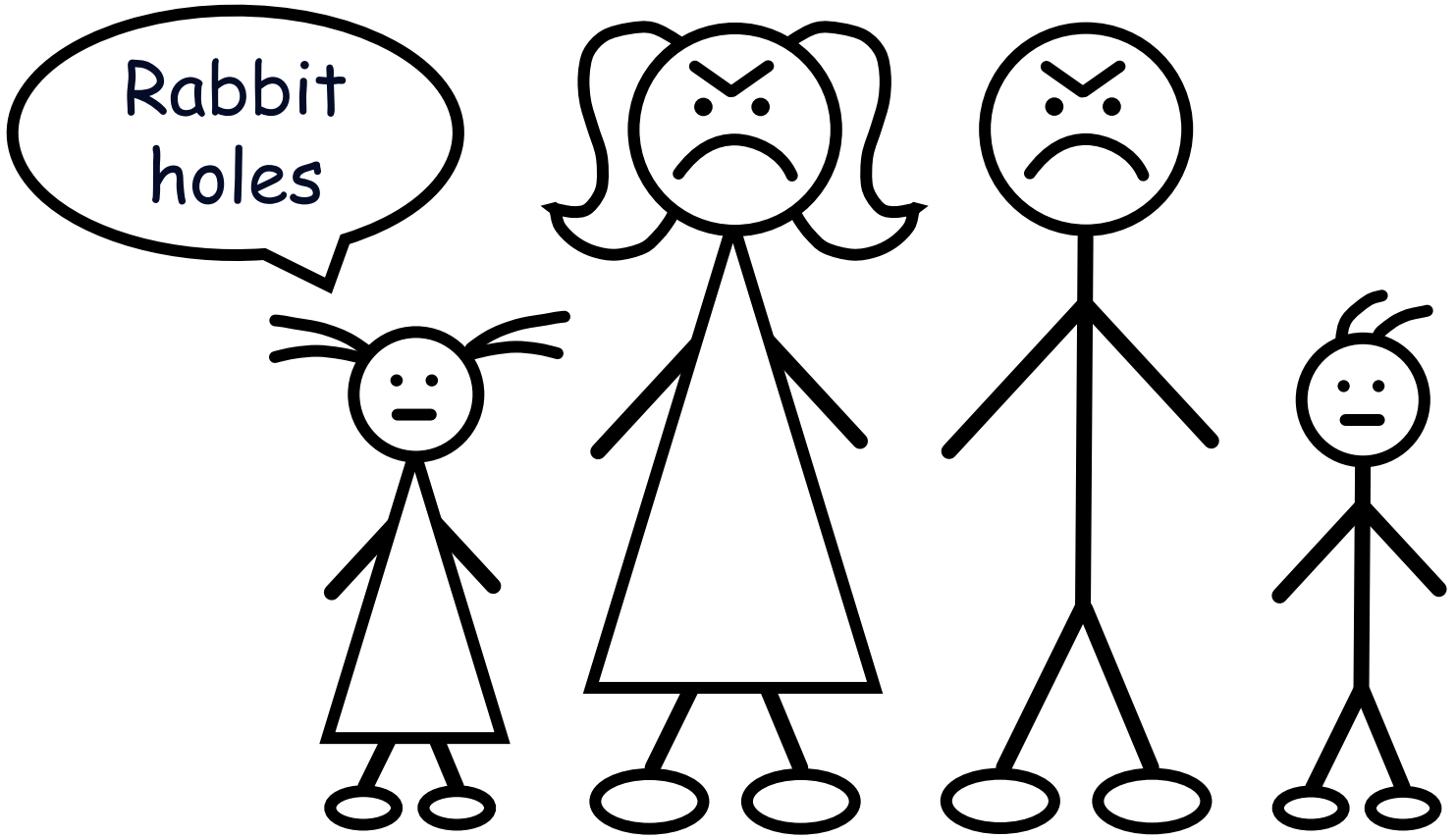
Talk to other people

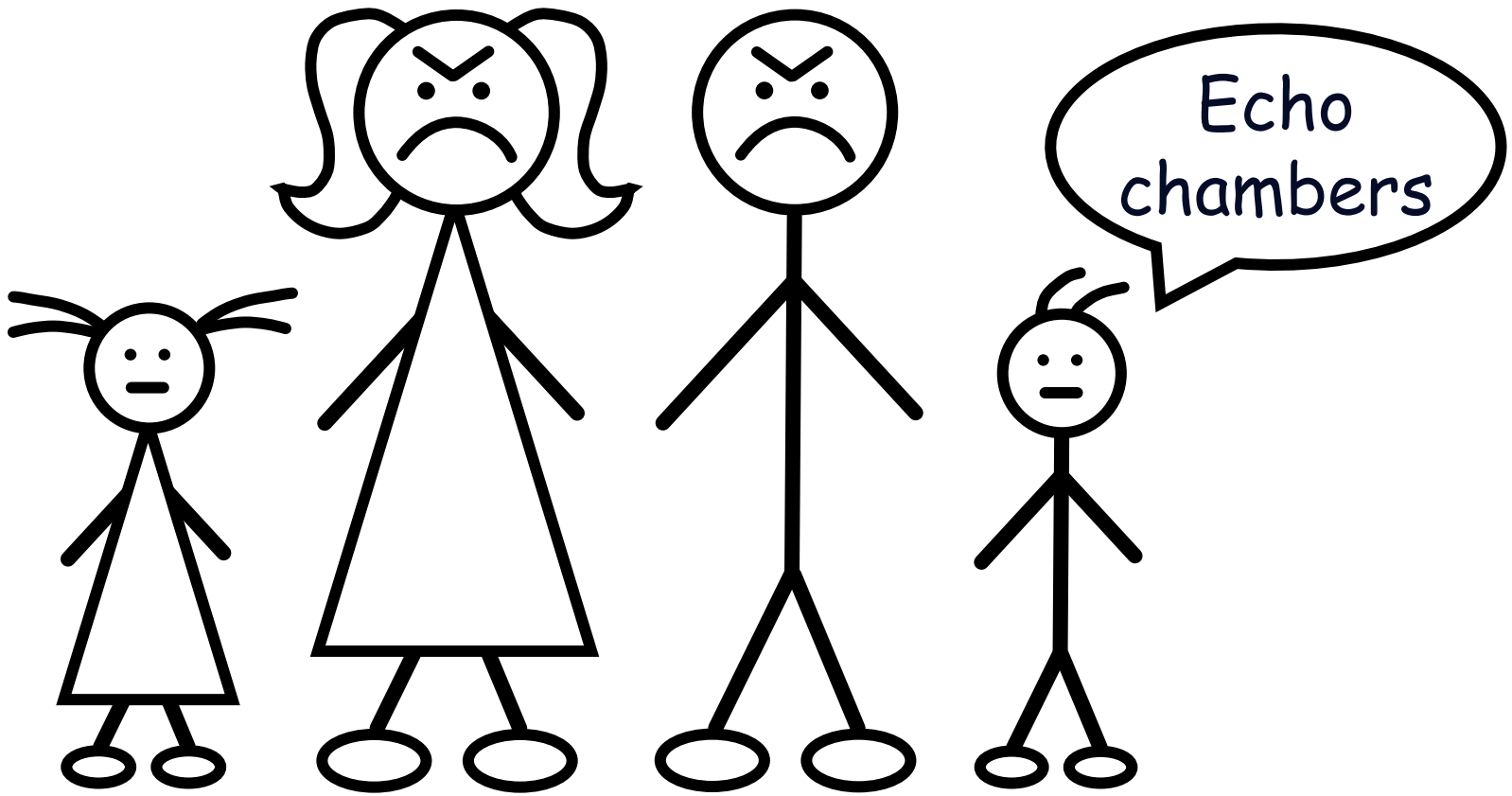
Alice and Bob in AI Land

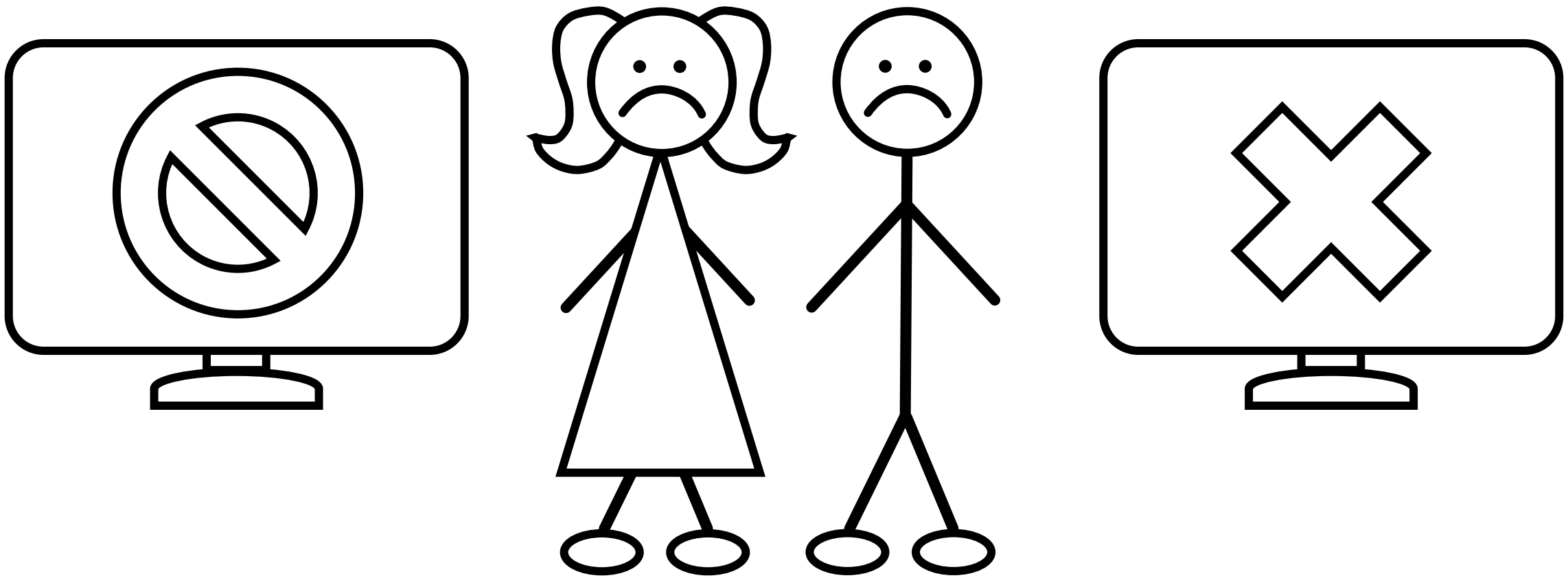
(continued …)

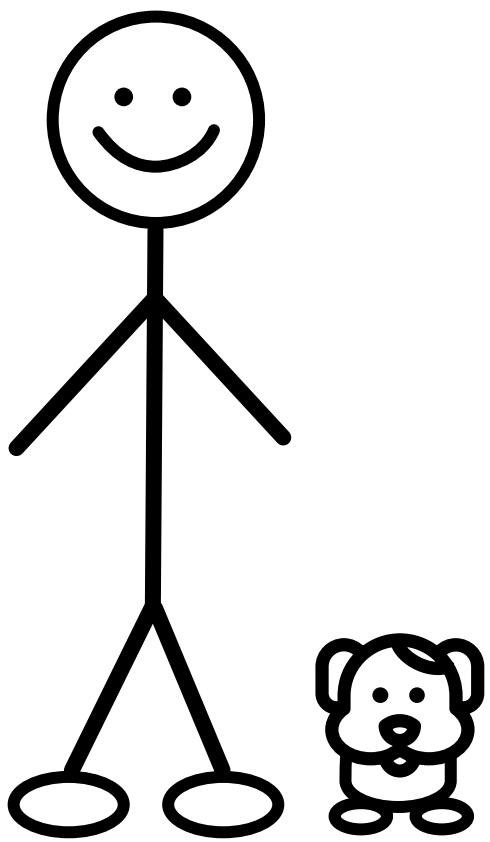


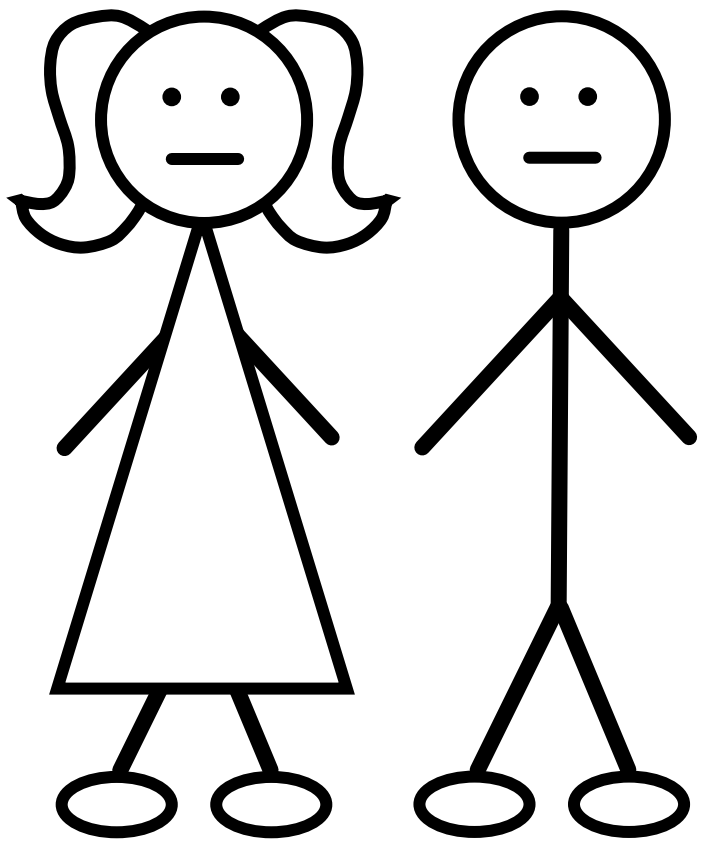

Recommendation
Algorithms 101

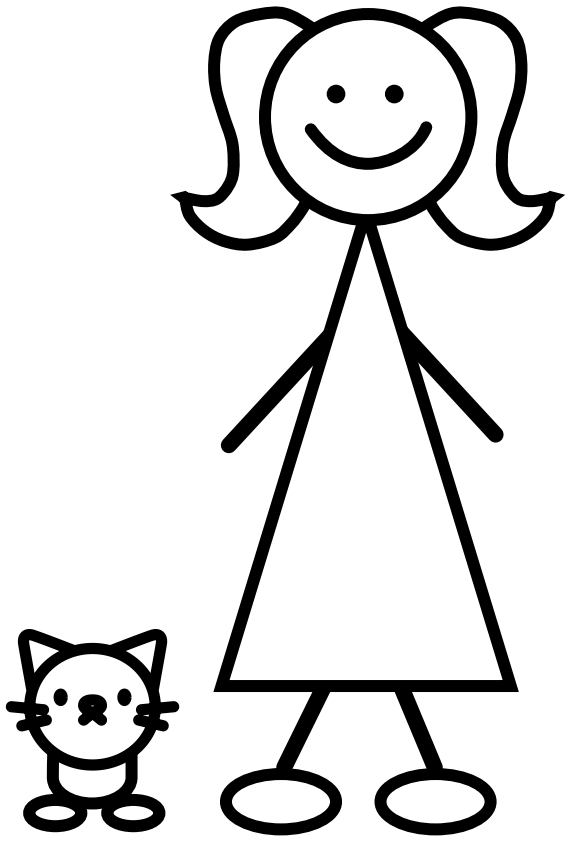

Rabbit
holes

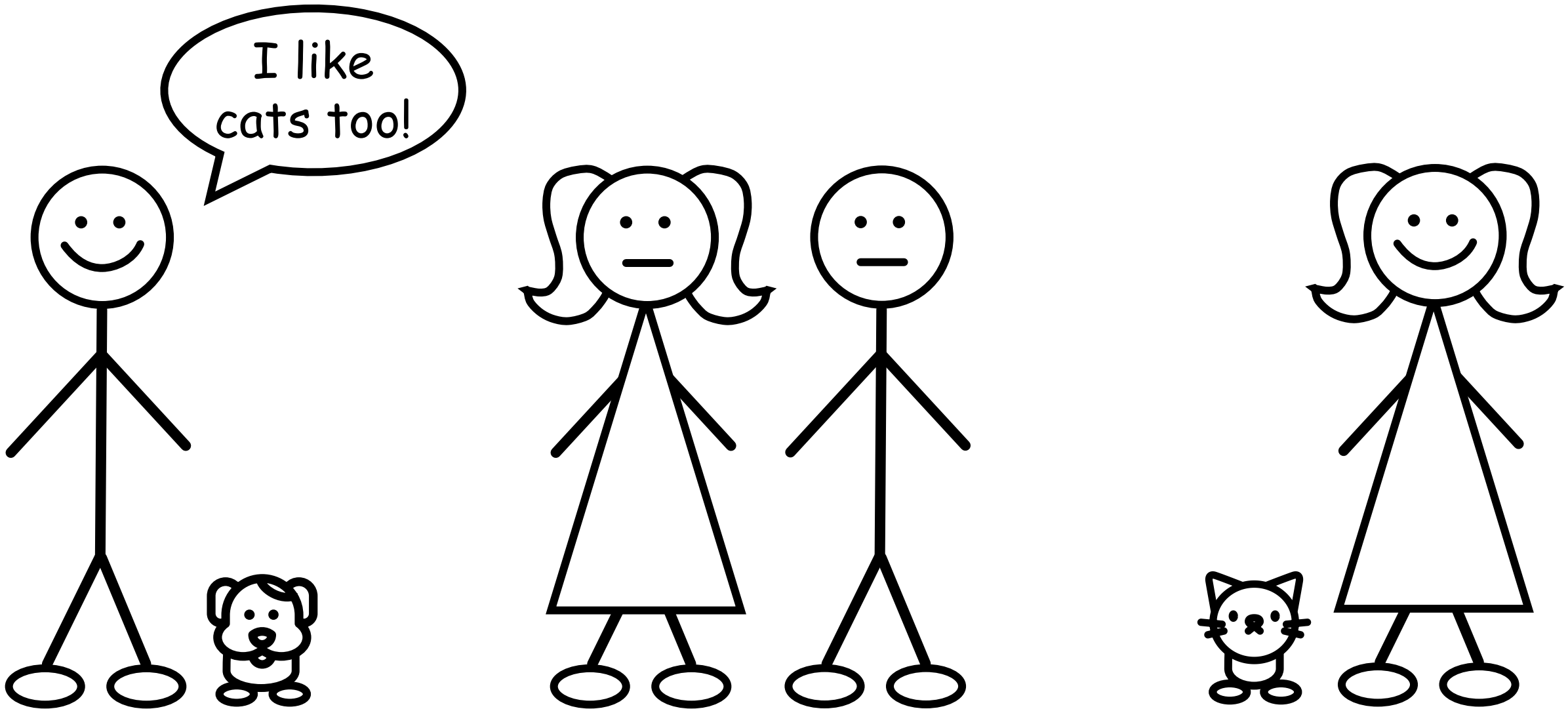

Echo
chambers

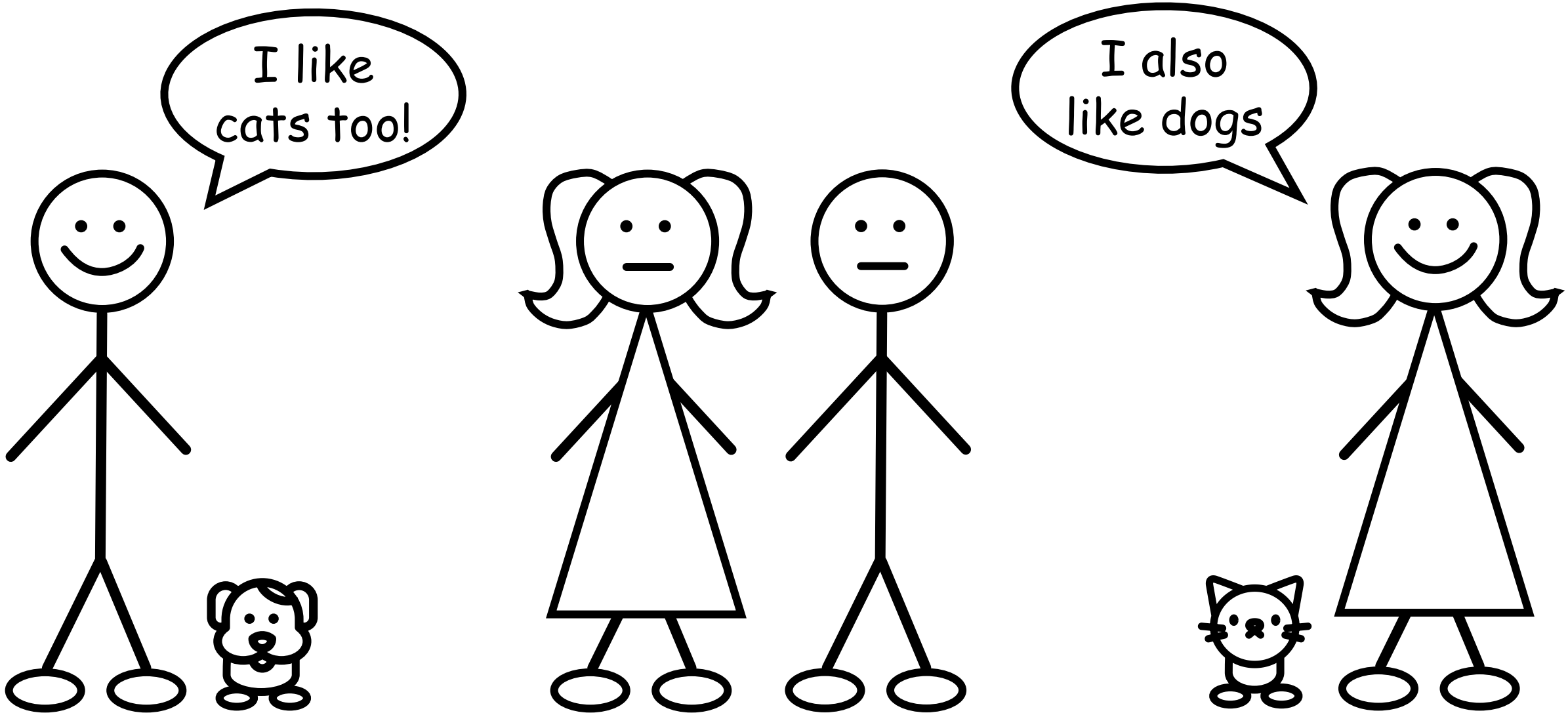


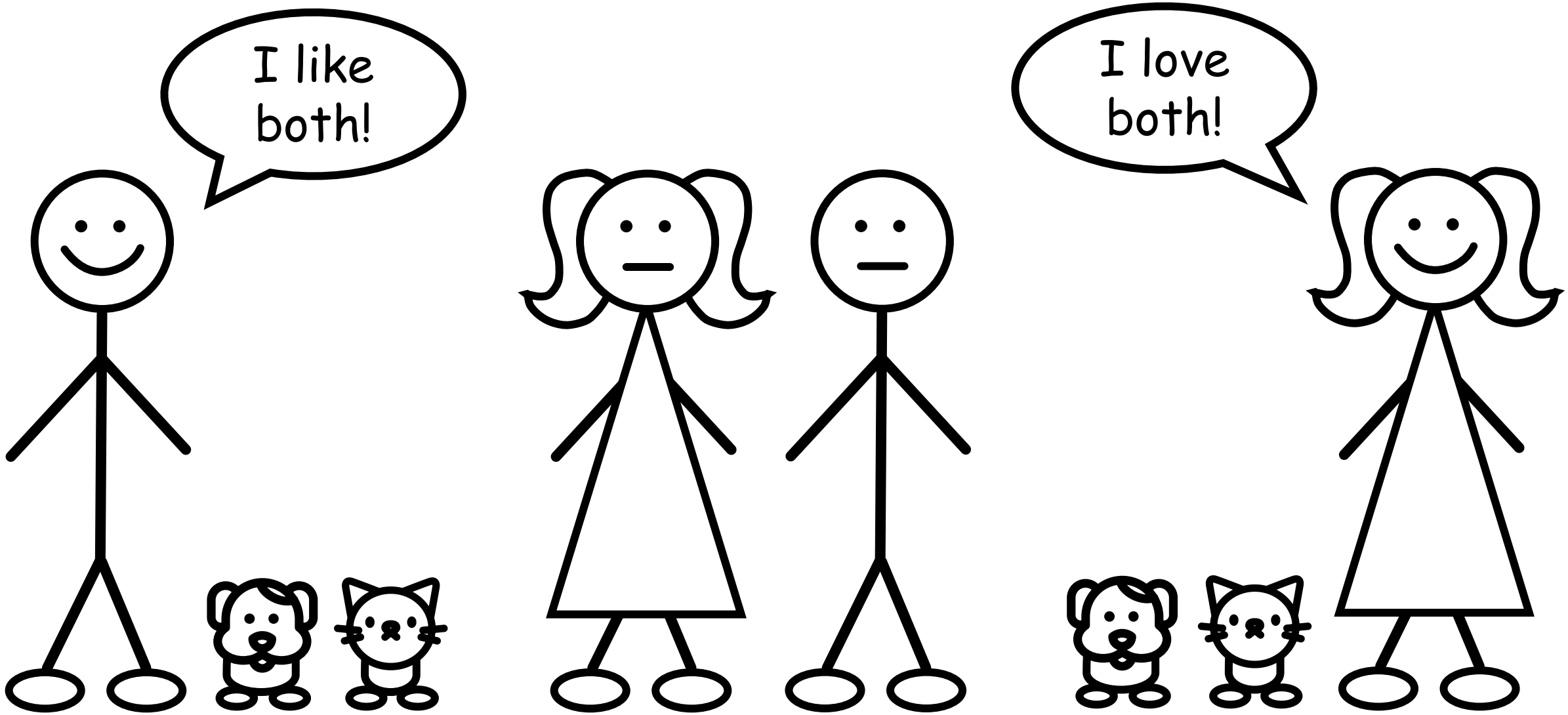


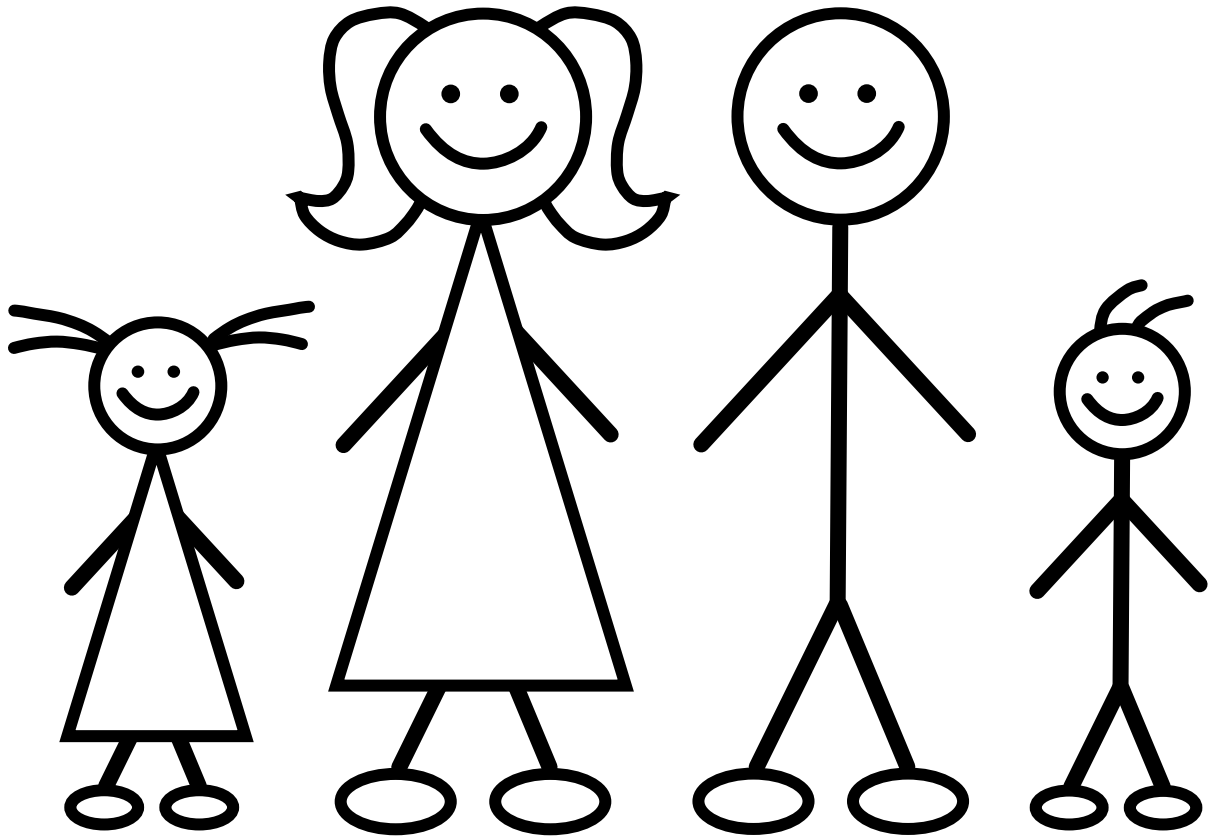


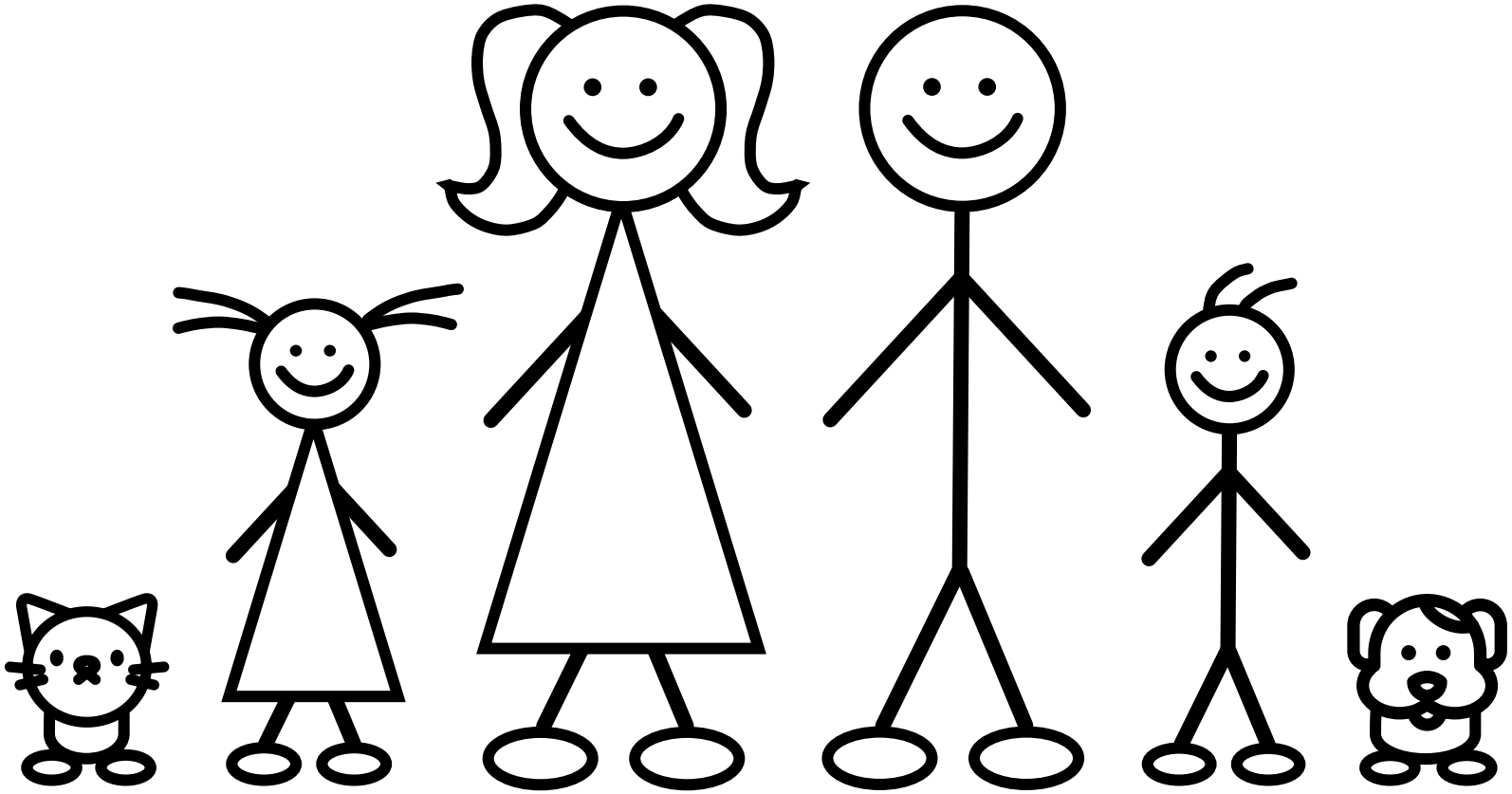


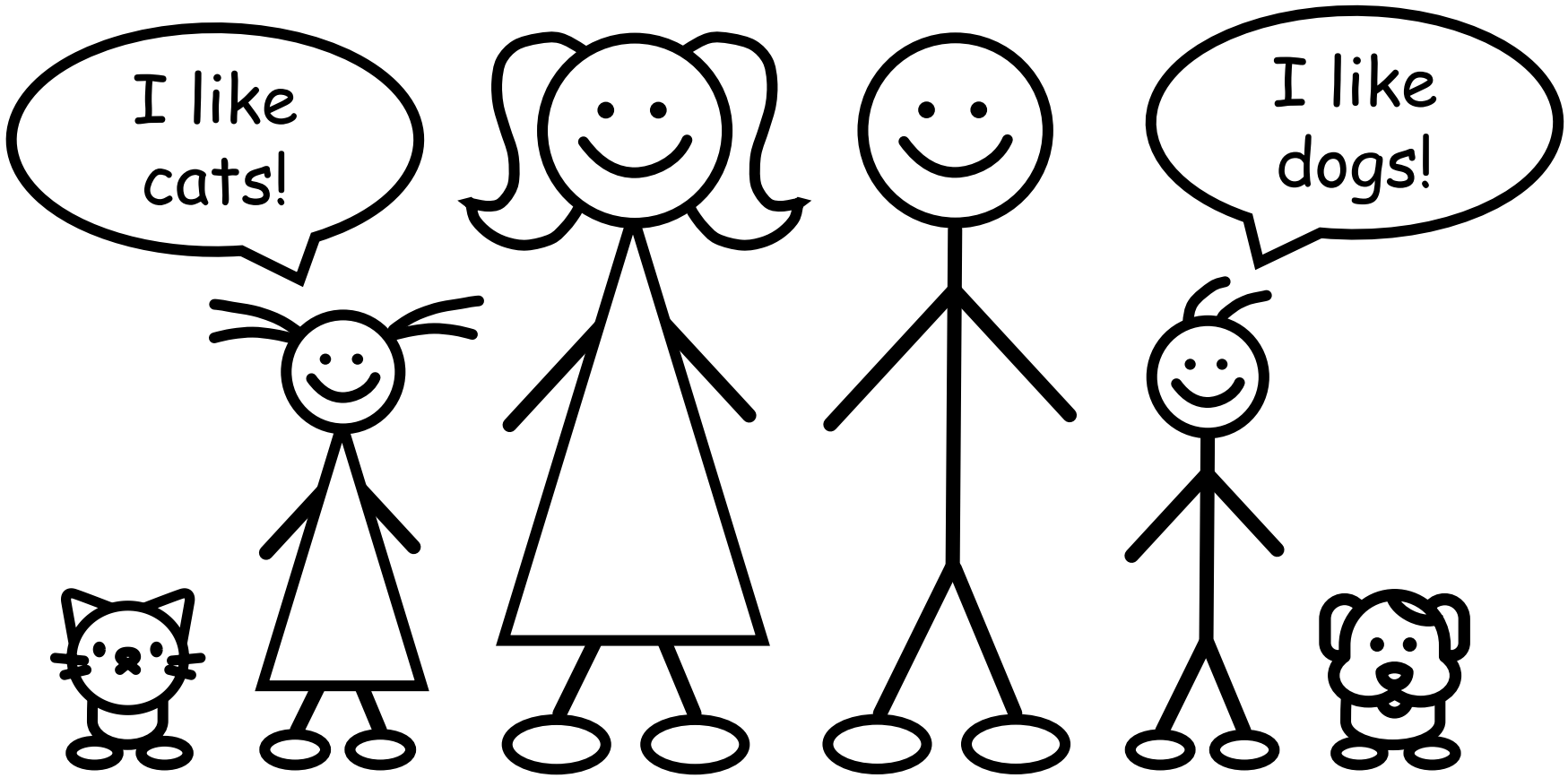

I like
cats too!


I like cats too!
I also like dogs

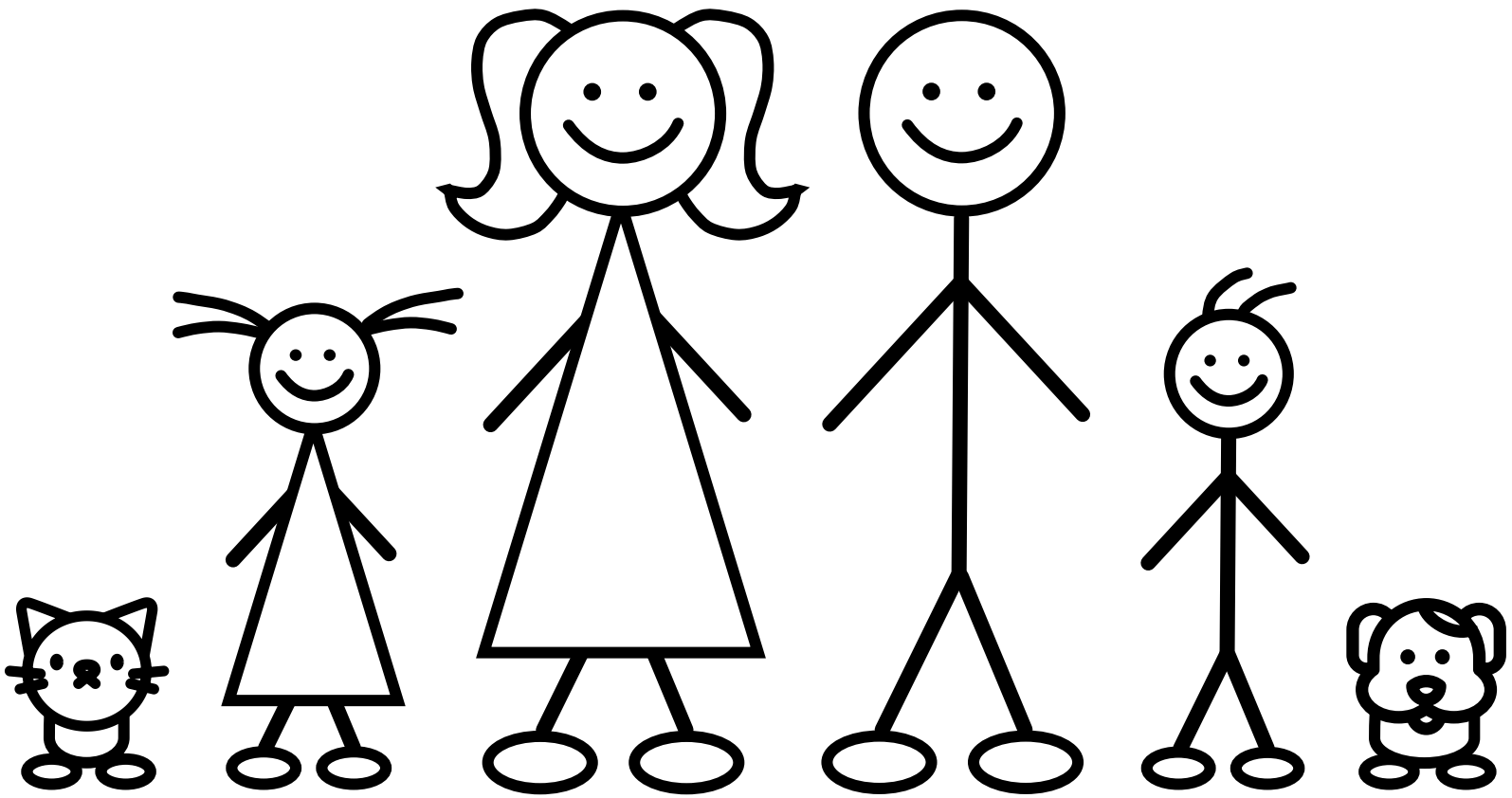

I like both!
I love both!






I like cats!
I like dogs!

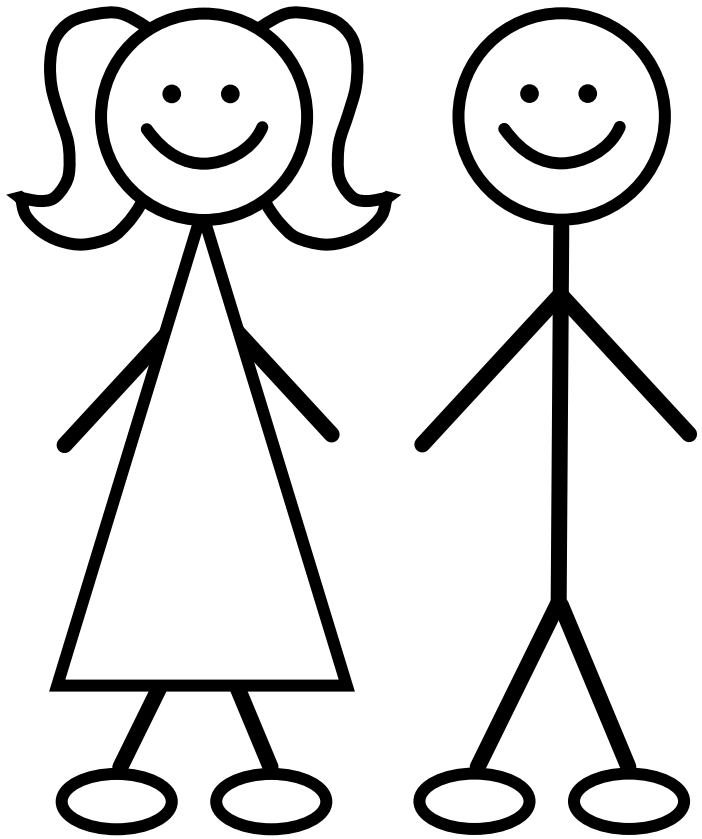
Happily ever after…

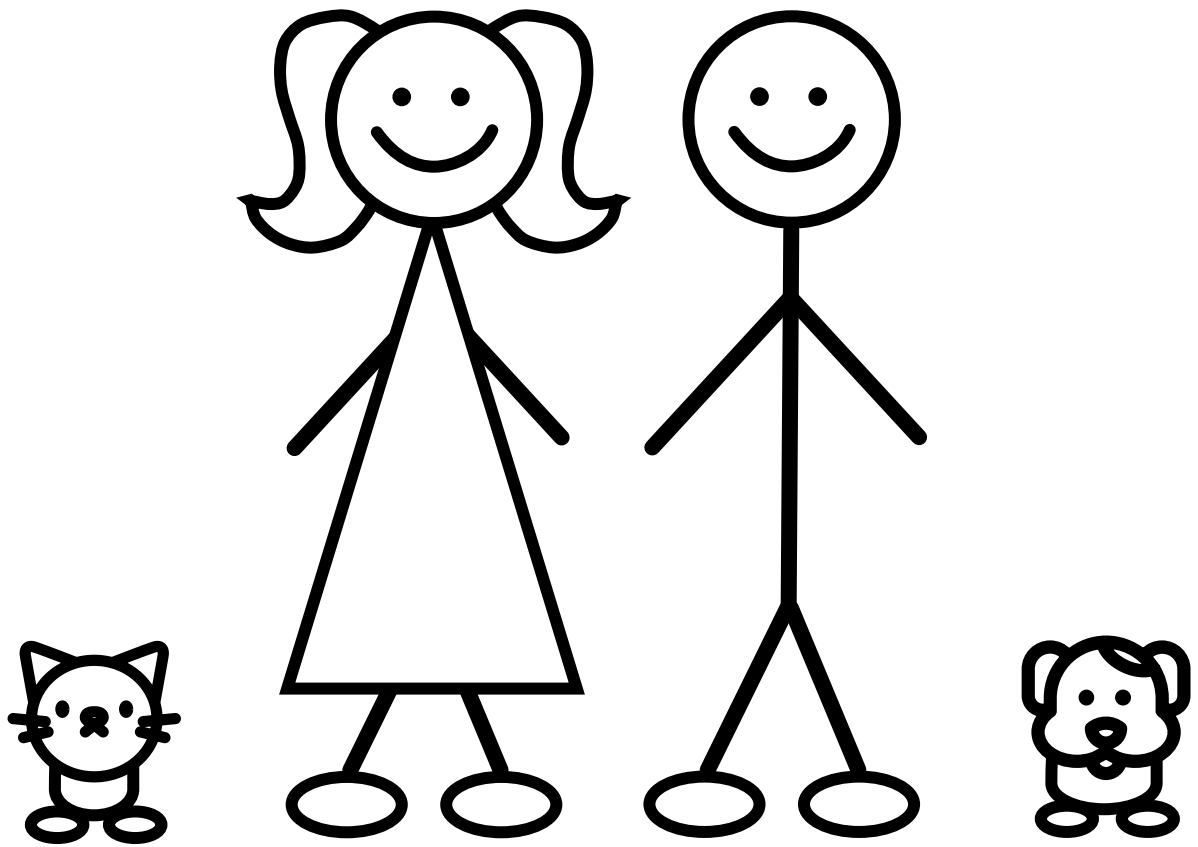


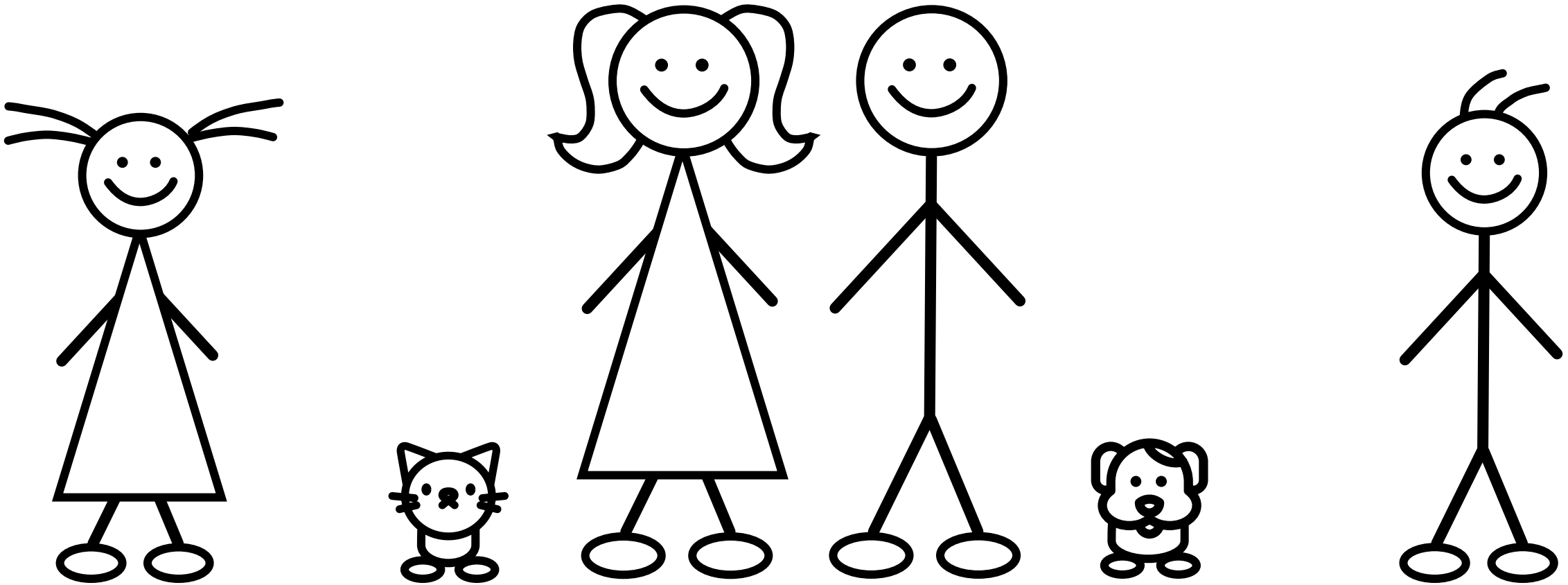
AI Ethics for Tomorrow

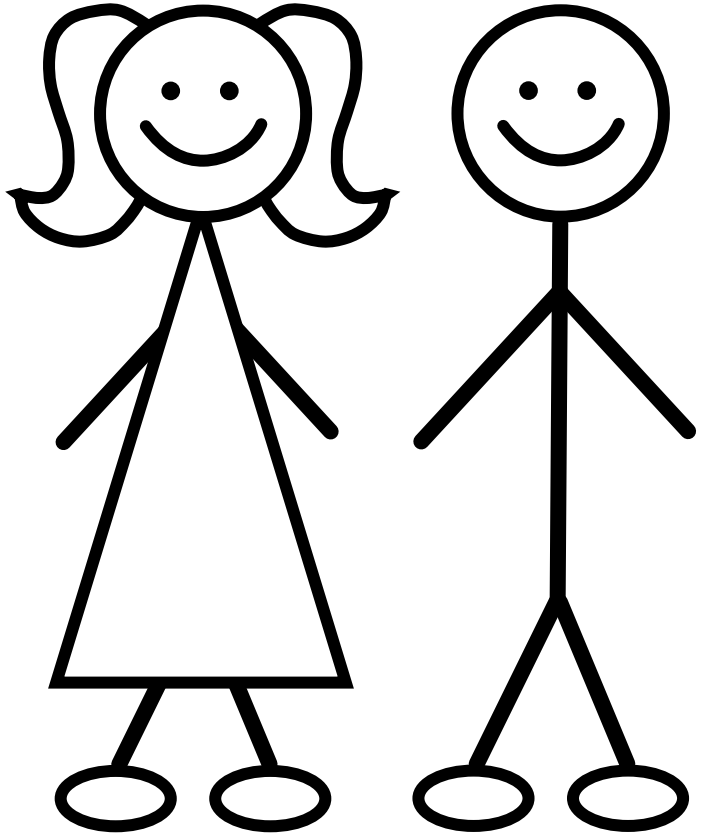
Alice and Bob in AI Land

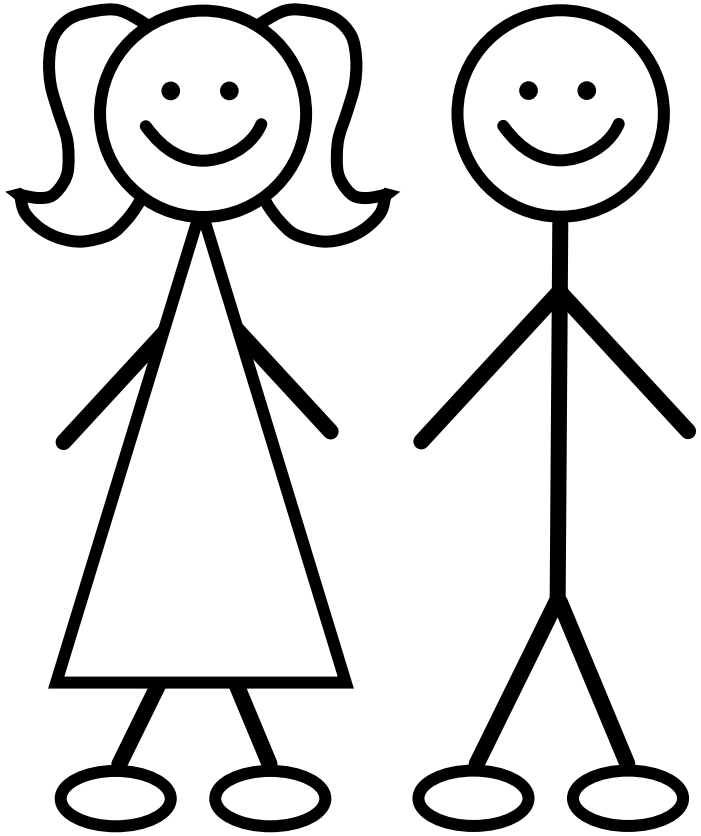
Chapter 2:
"AI Ethics of Tomorrow"

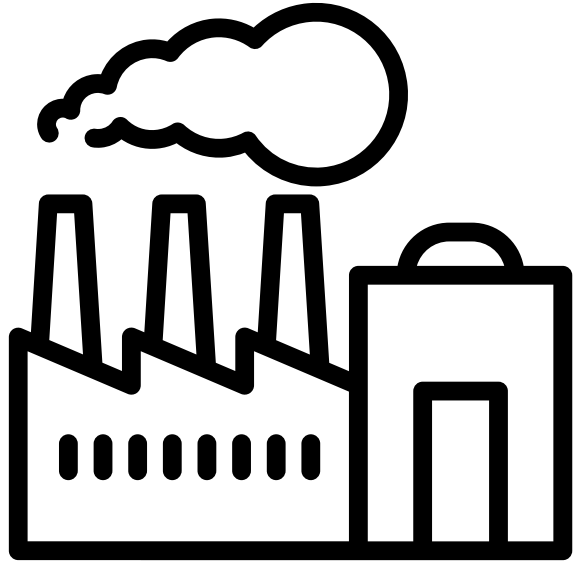


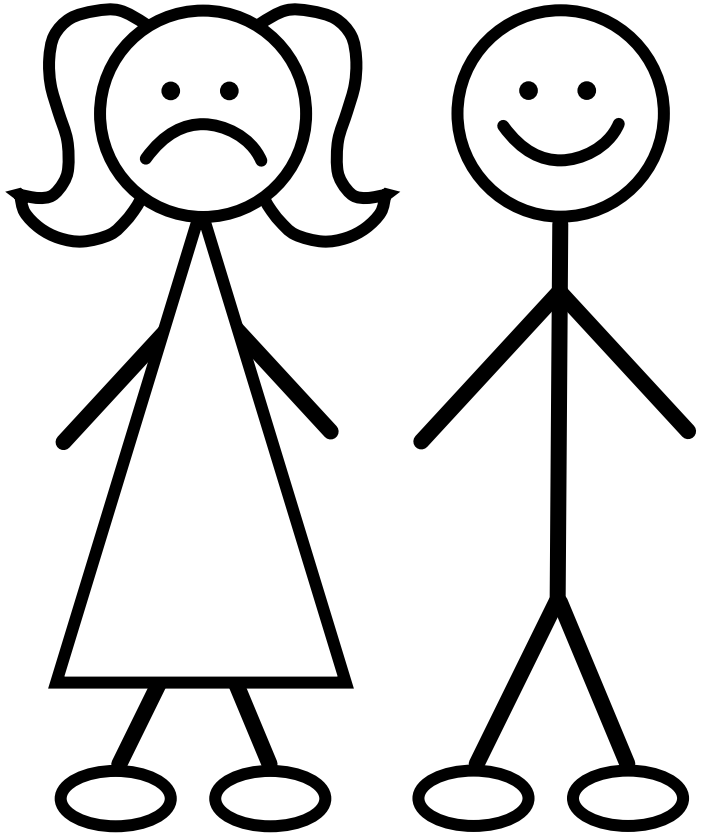


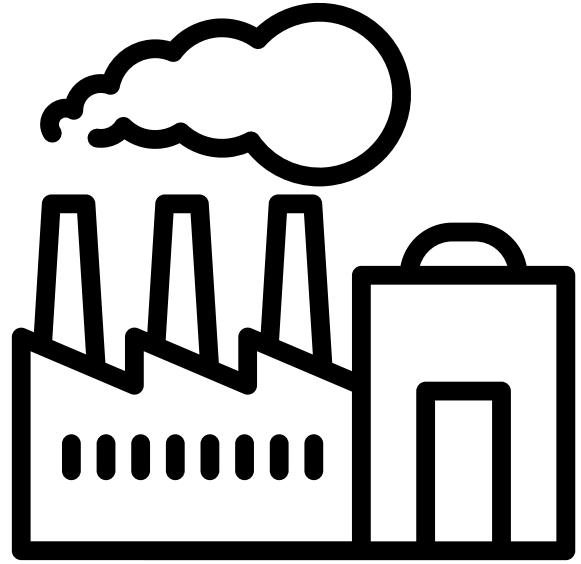


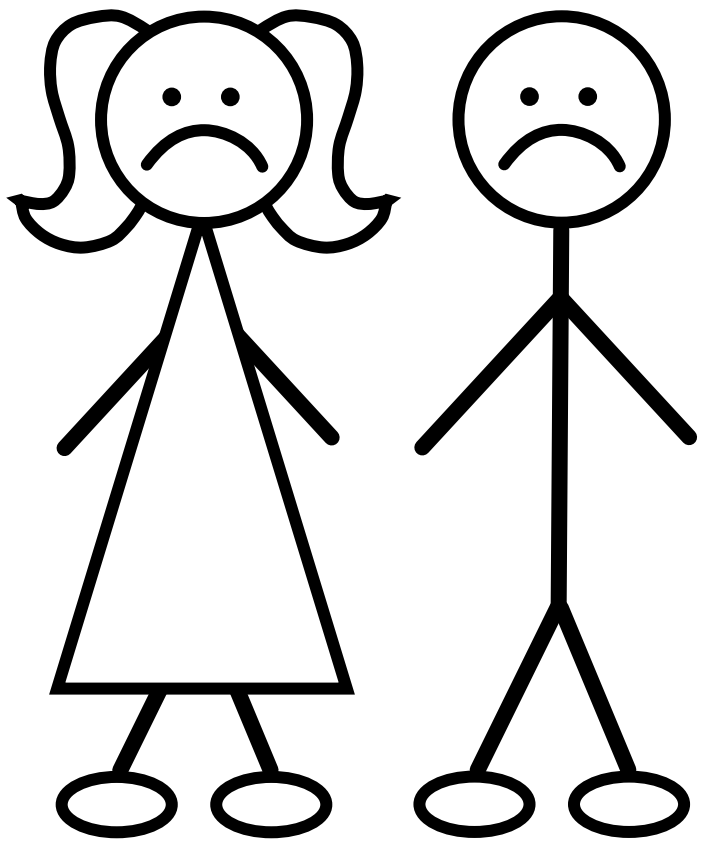


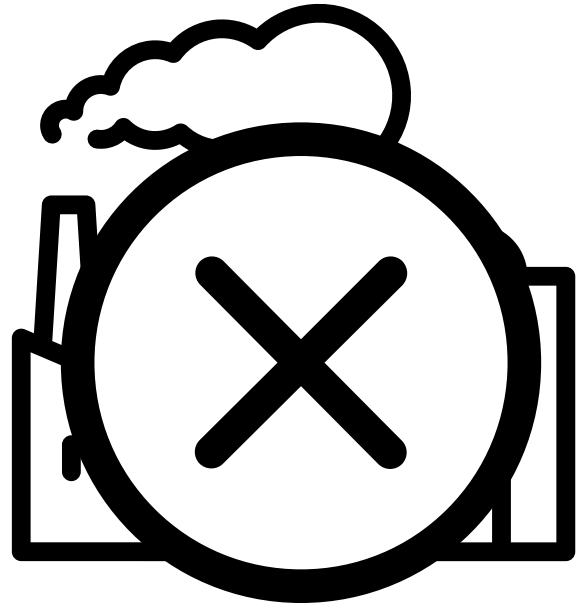


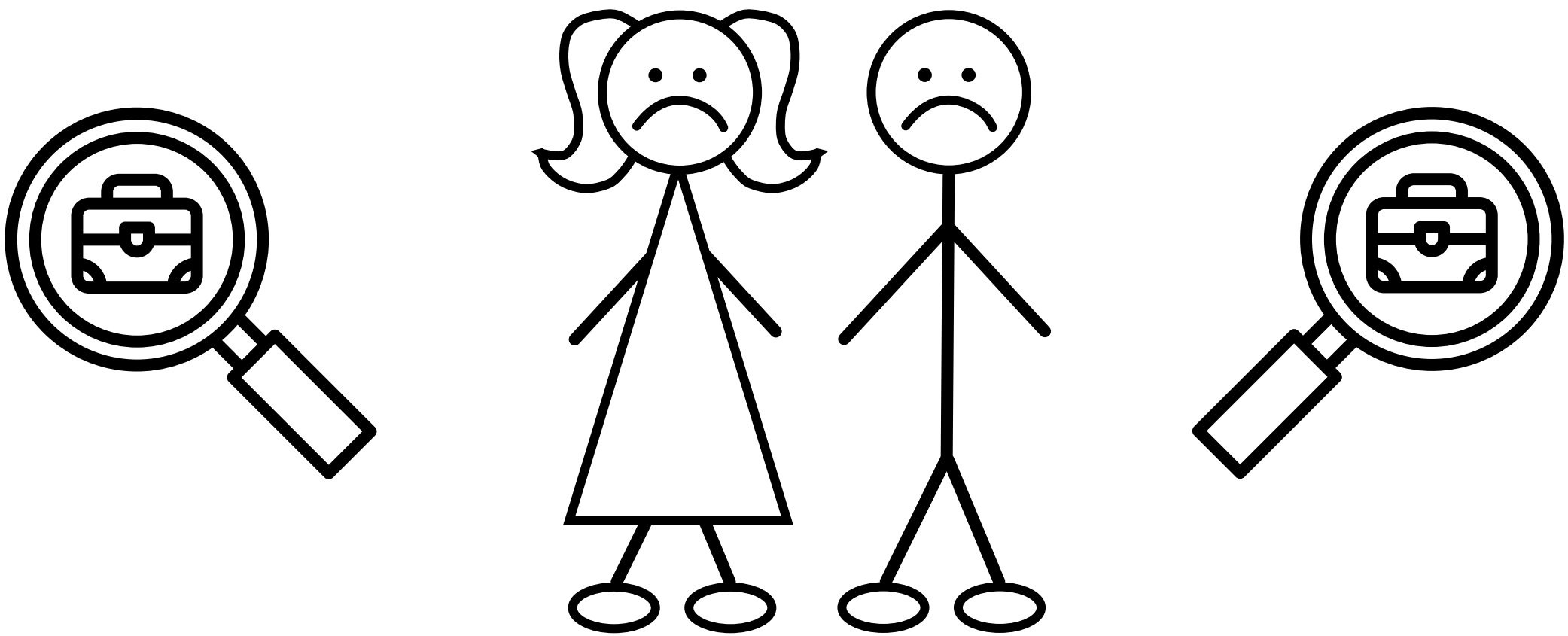


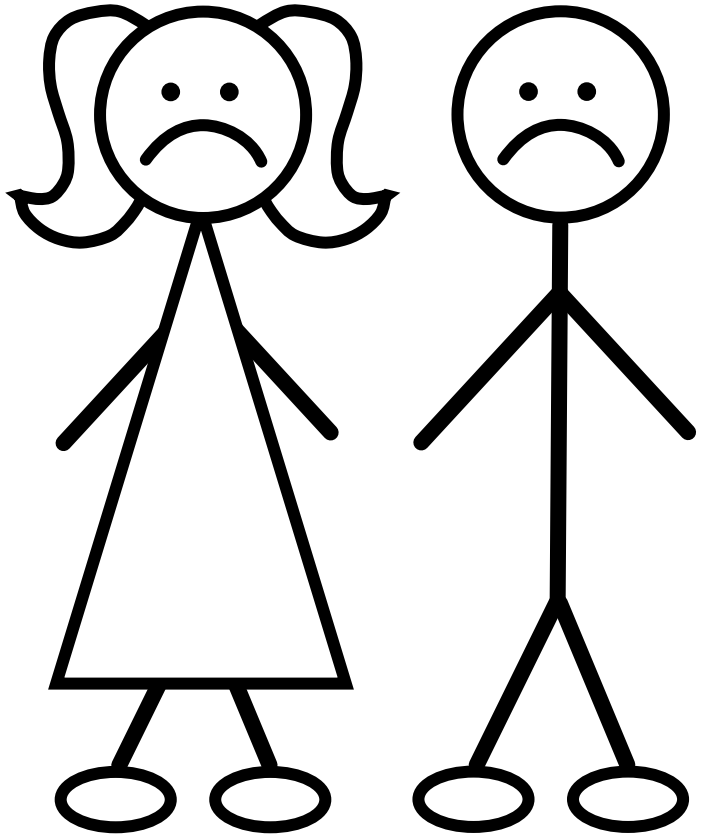


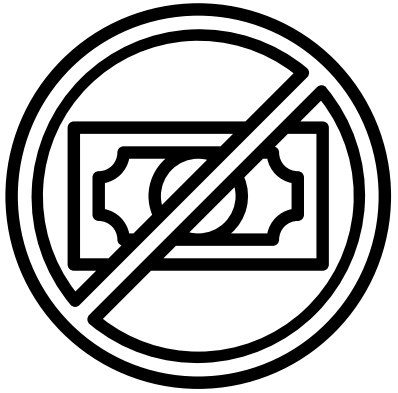


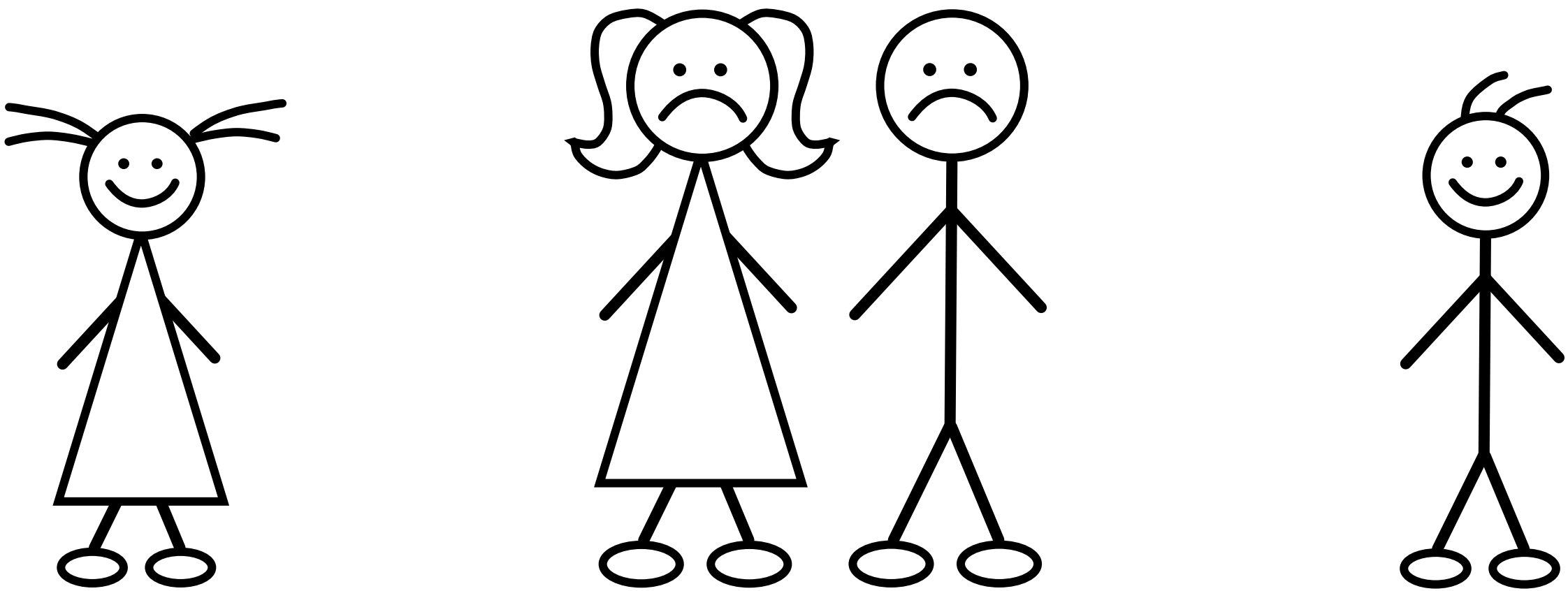


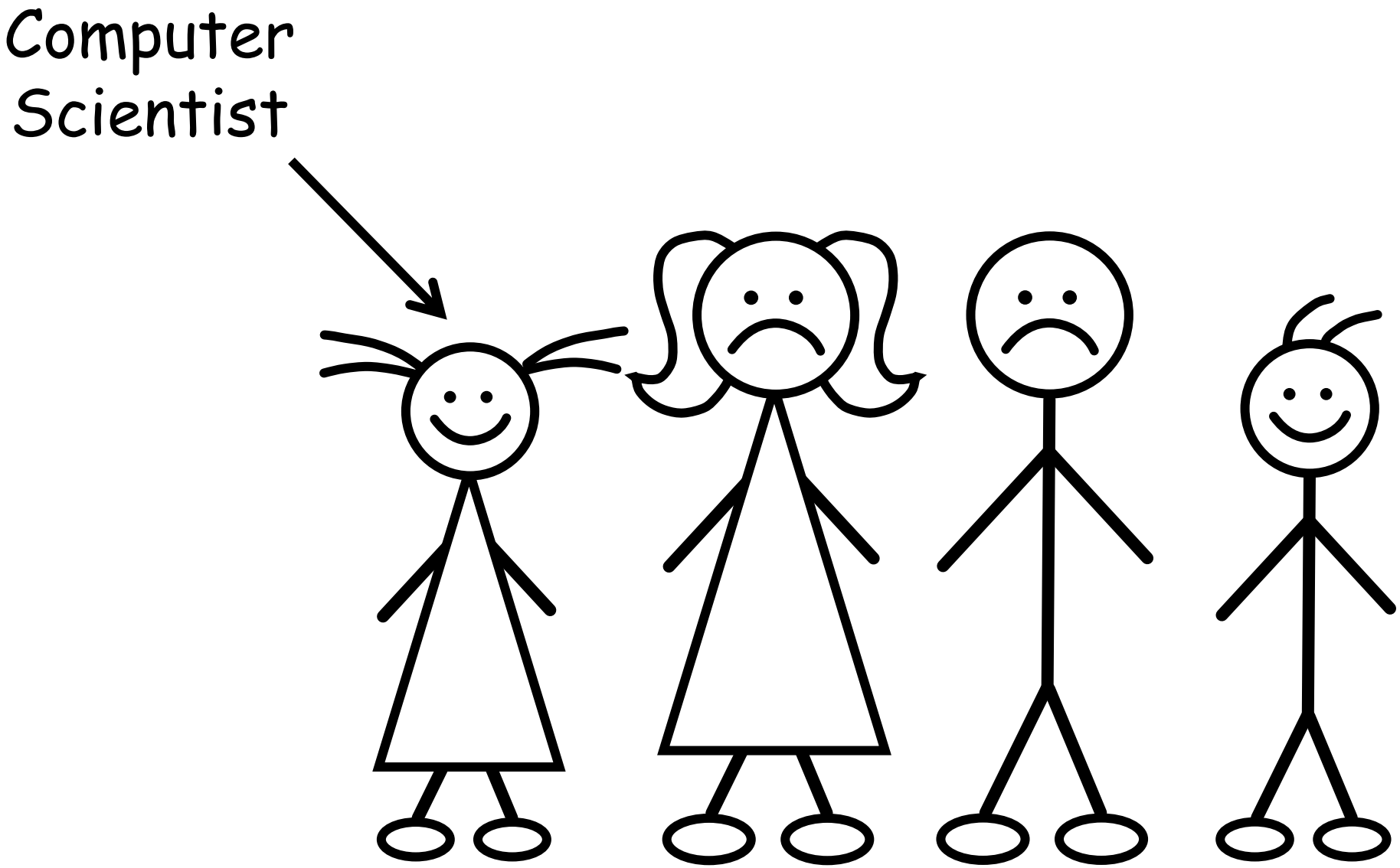


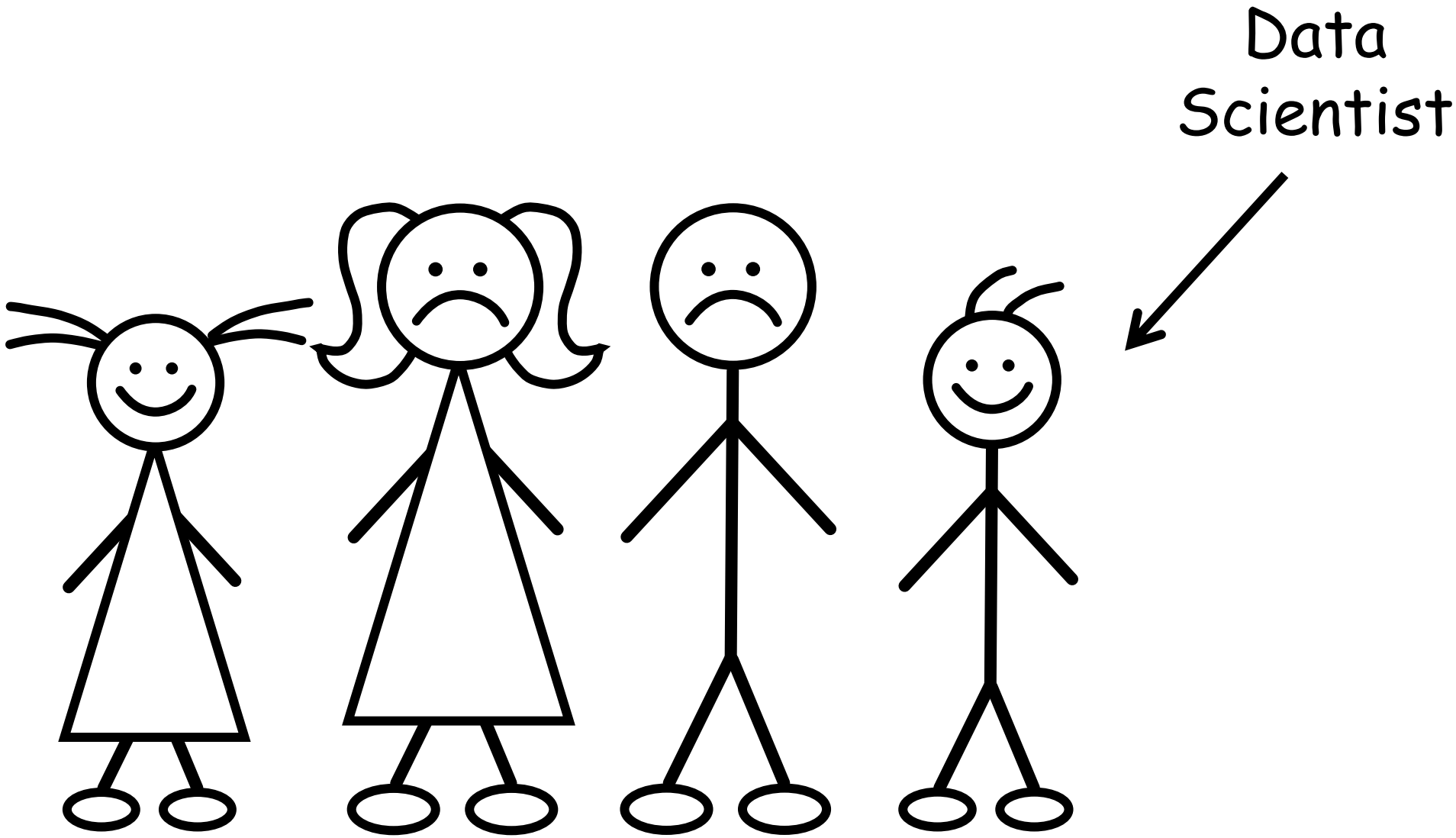


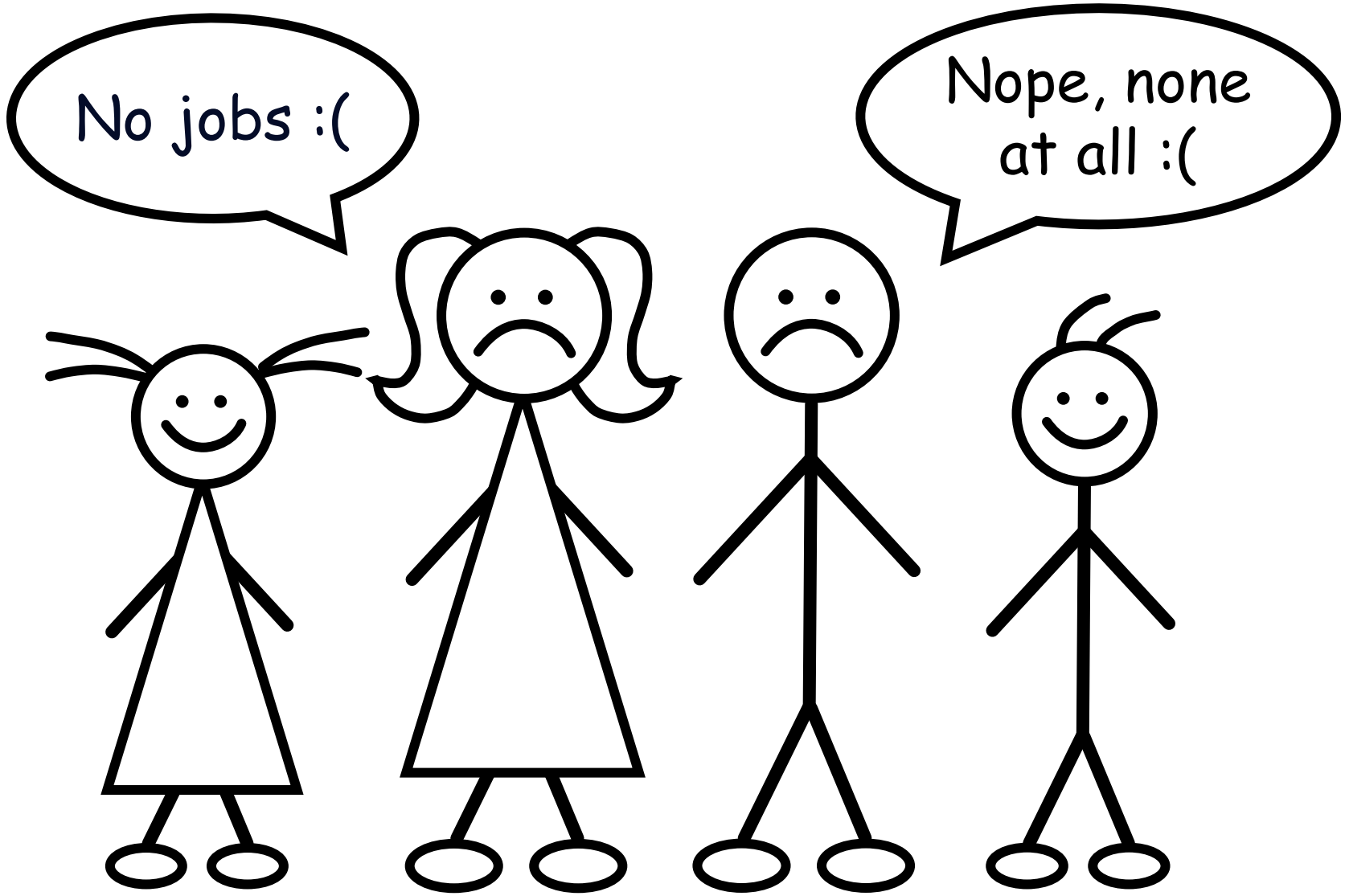


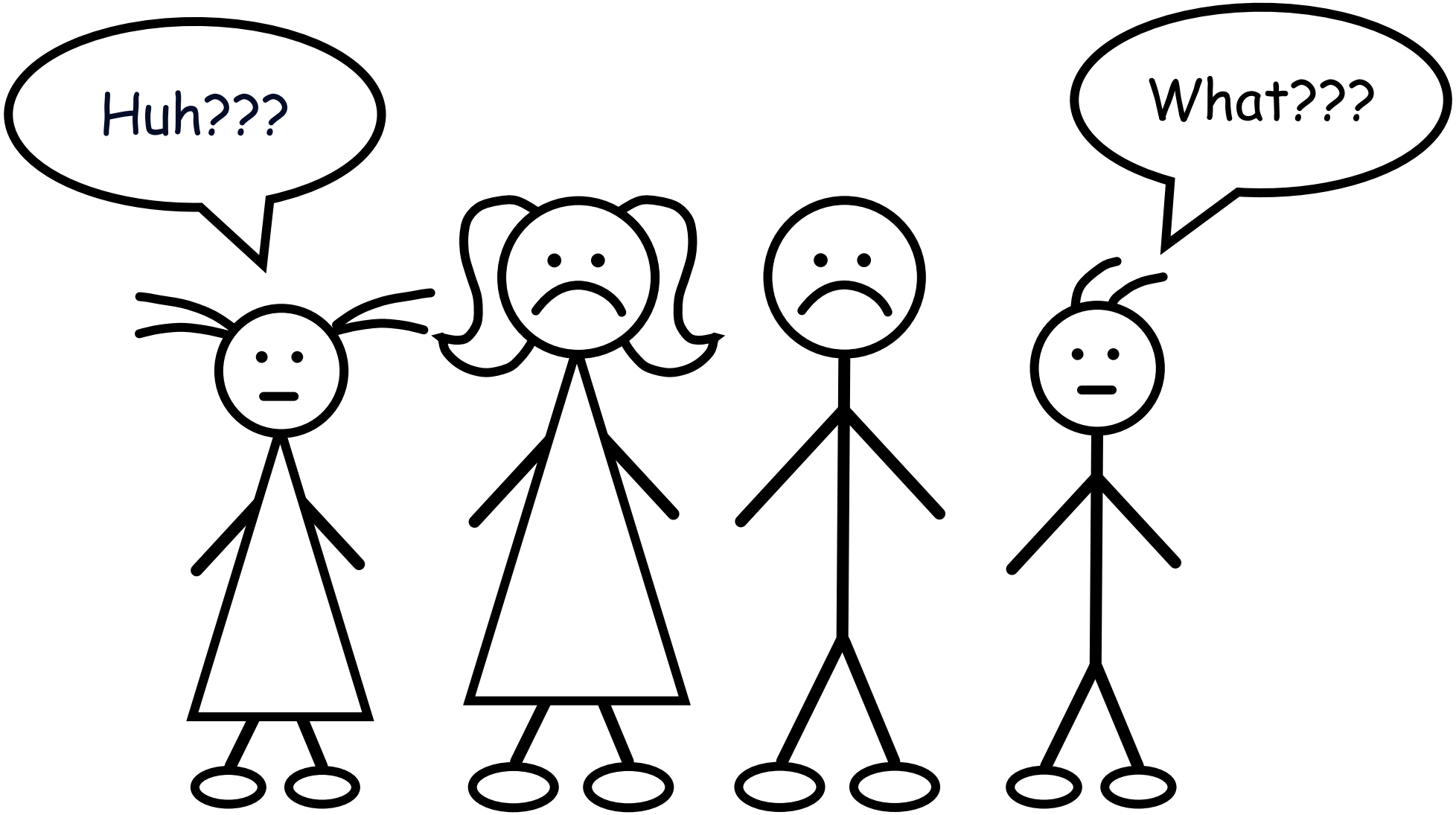


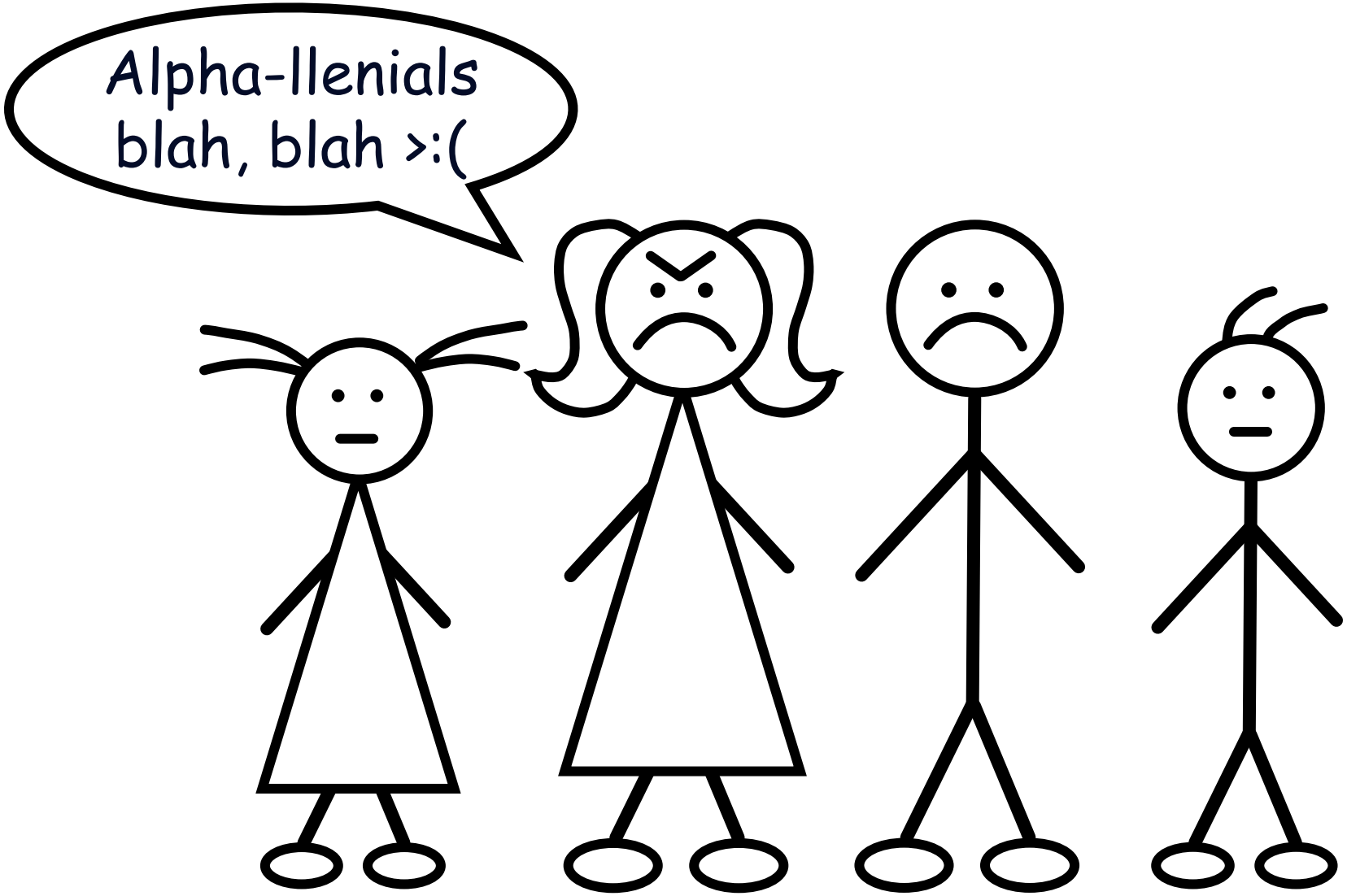


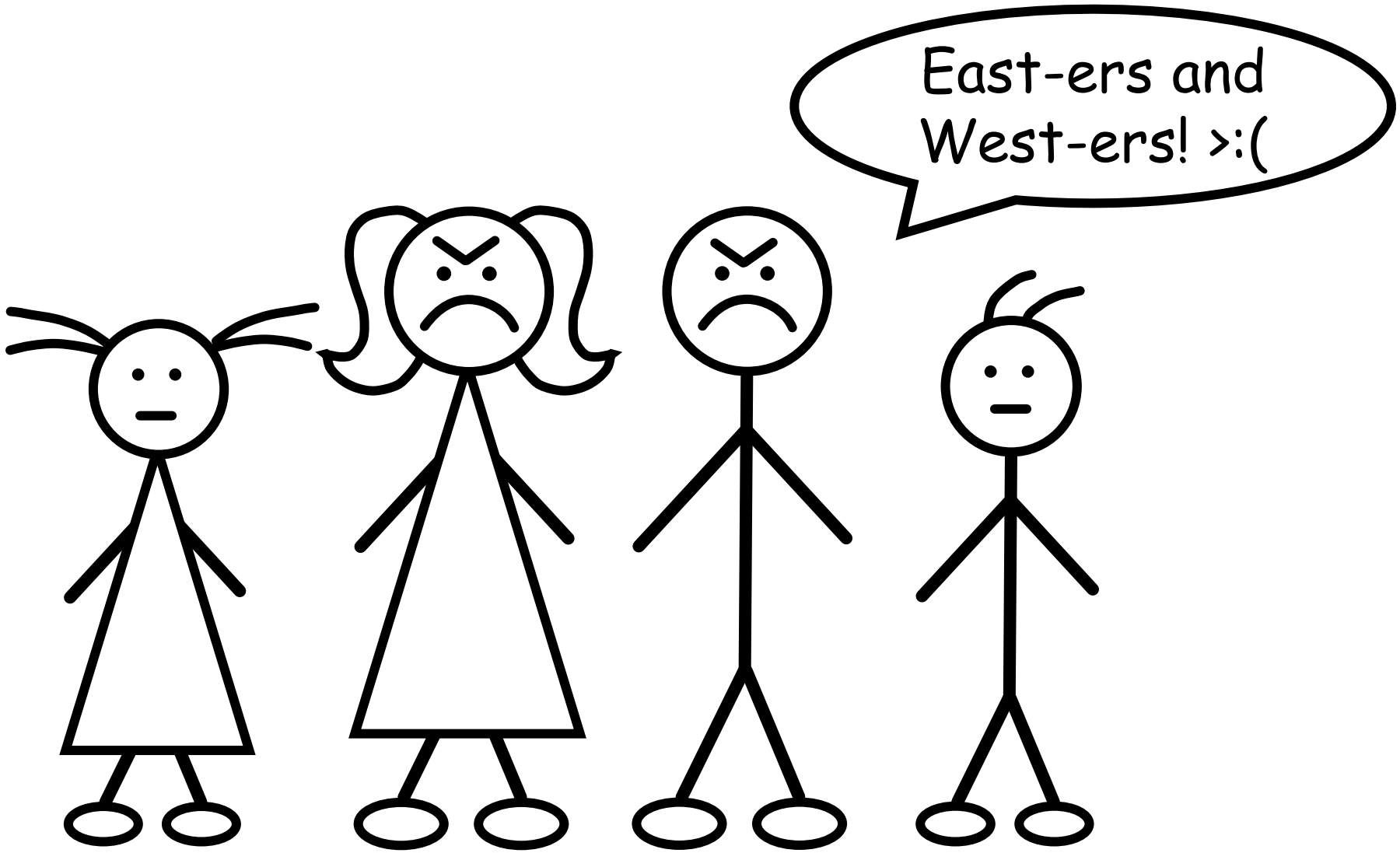

Computer
Scientist

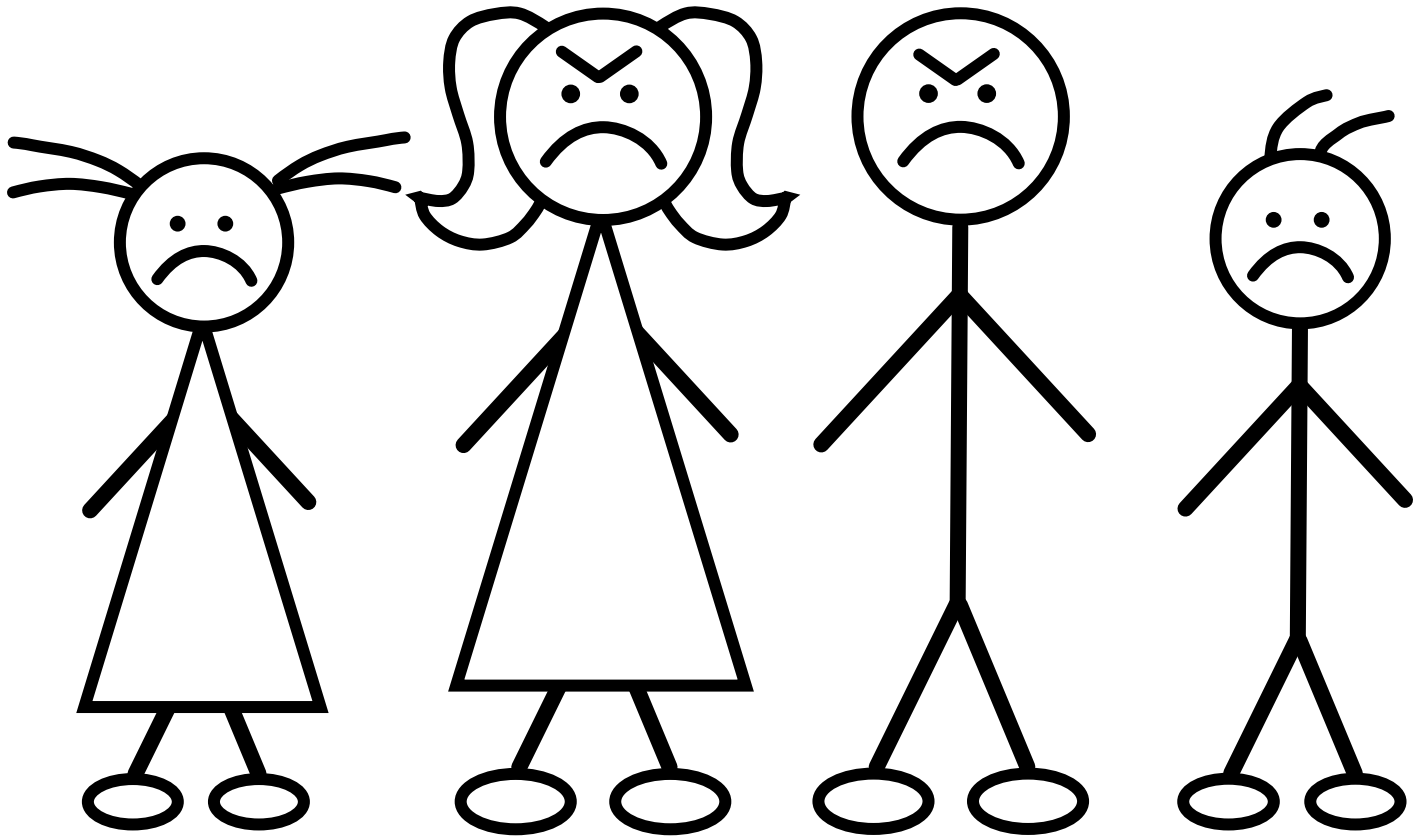

Data
Scientist

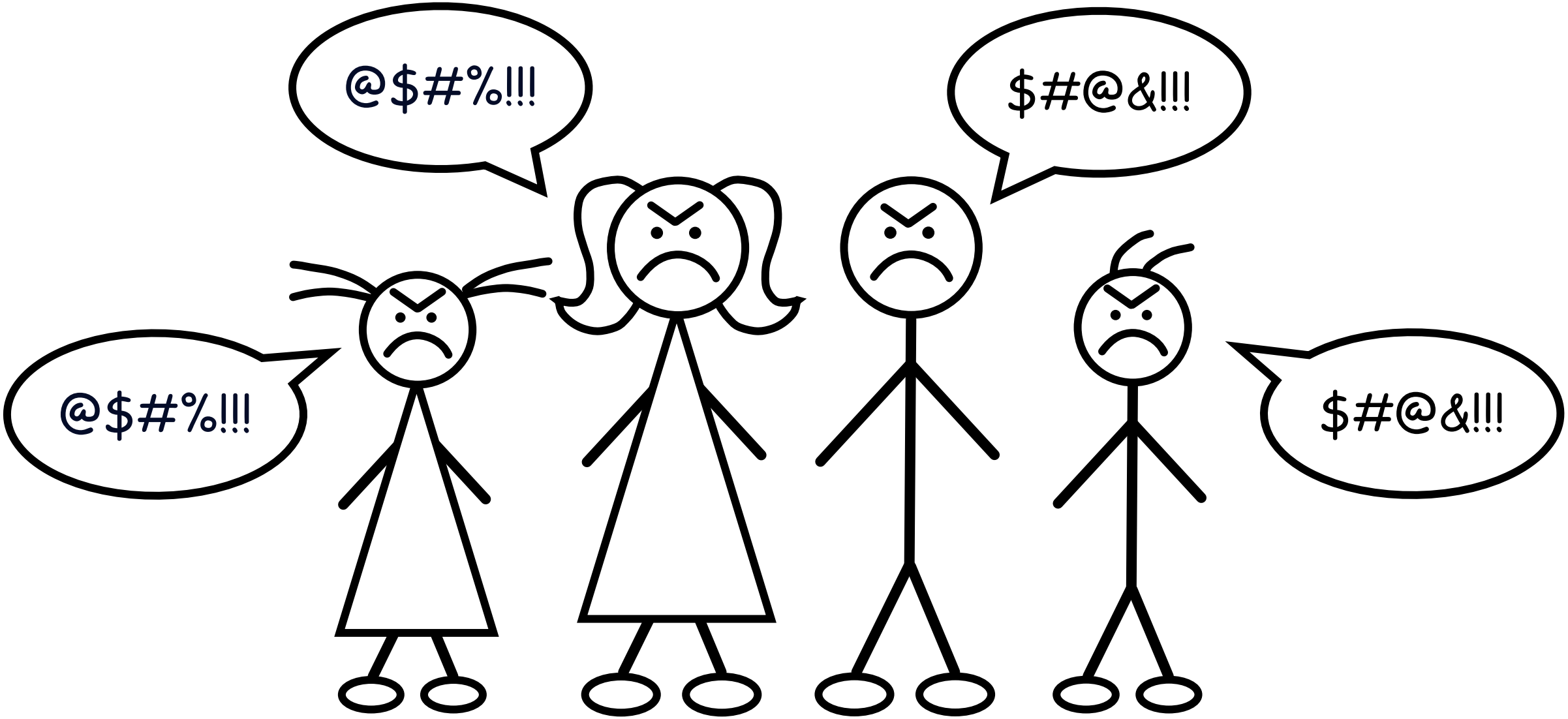

No jobs :(
Nope, none at all :(

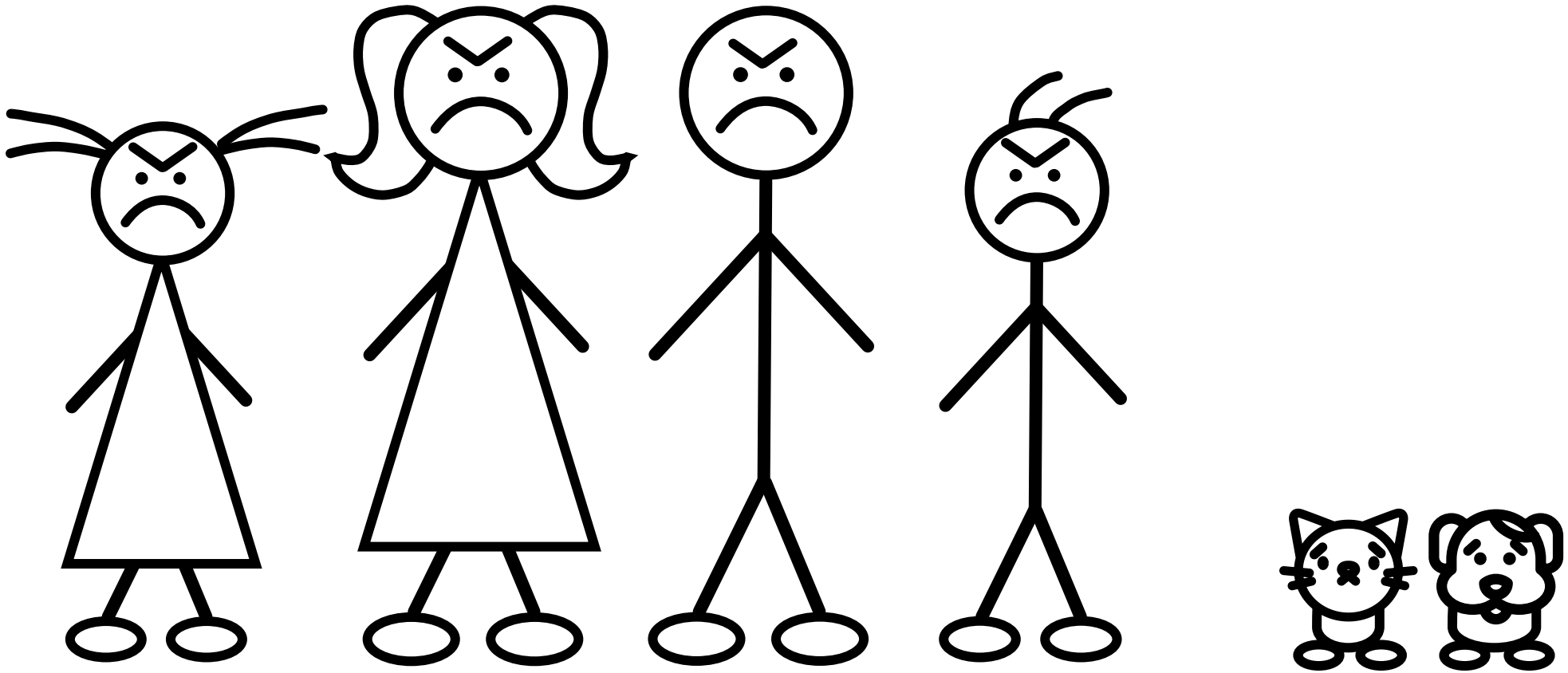

Huh???
What???


Alpha-llenials
blah, blah >:(


East-ers and
West-ers! >:(



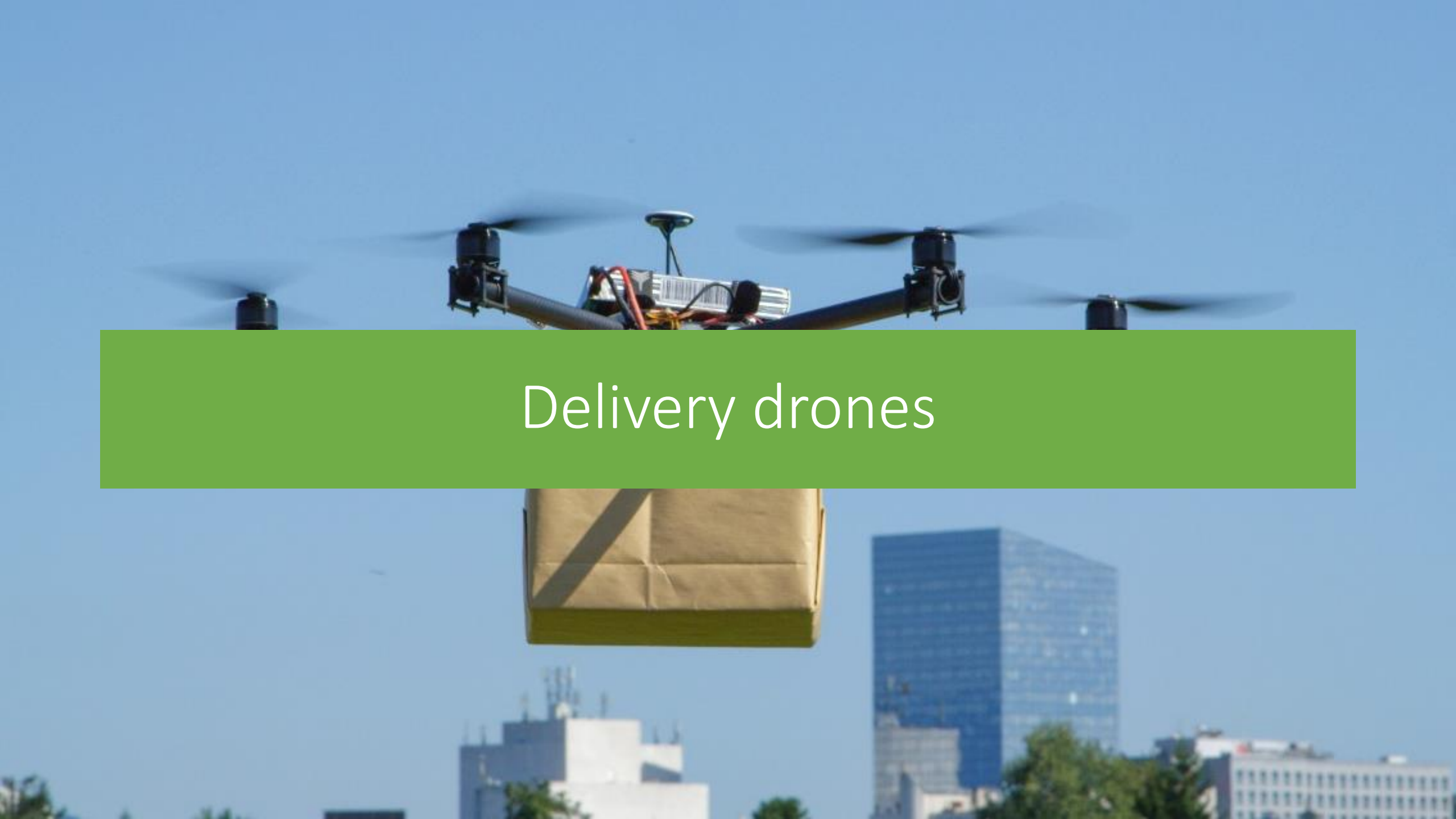




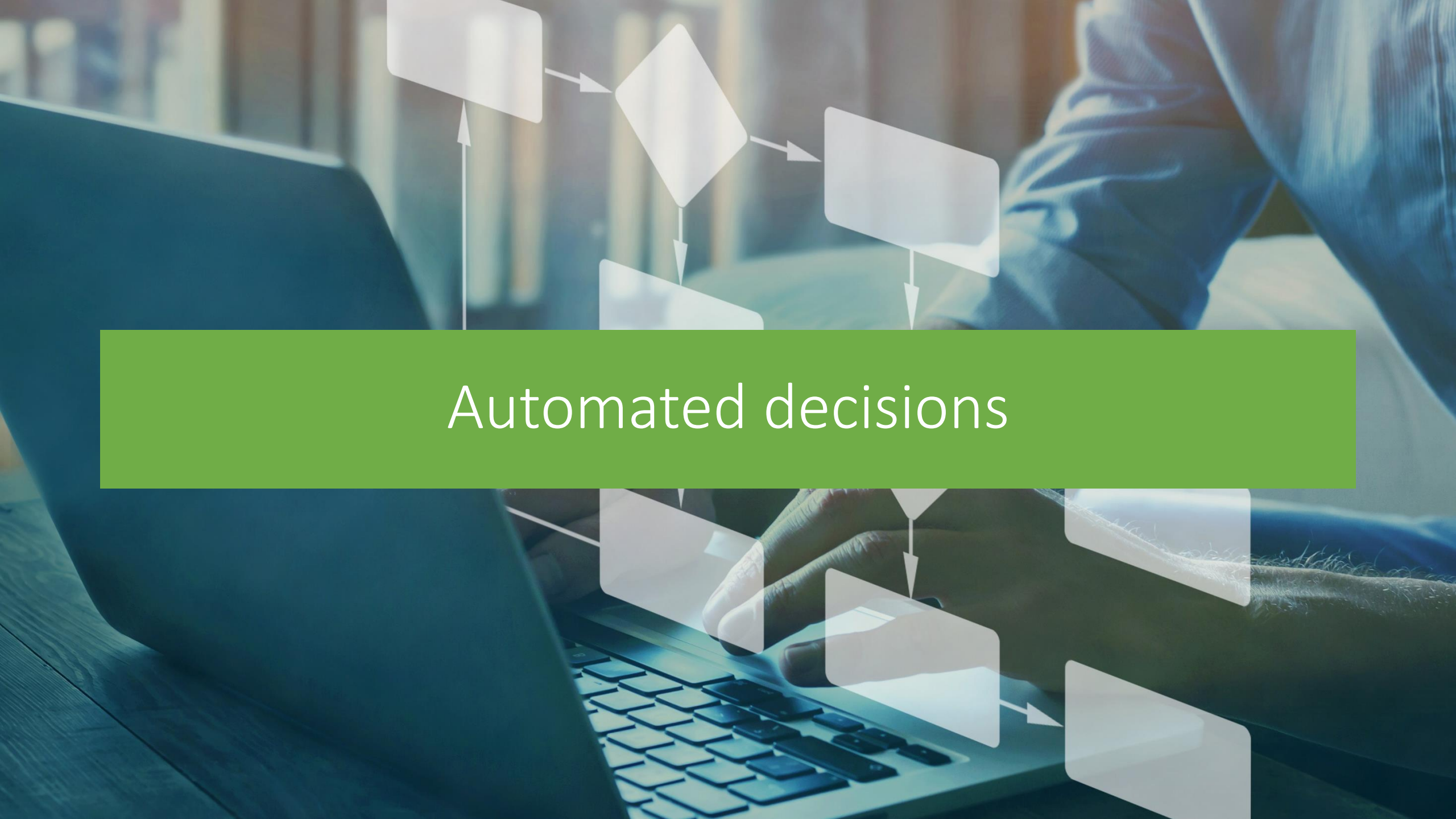
Benefits

Automated labor

Autonomous vehicles



Delivery drones

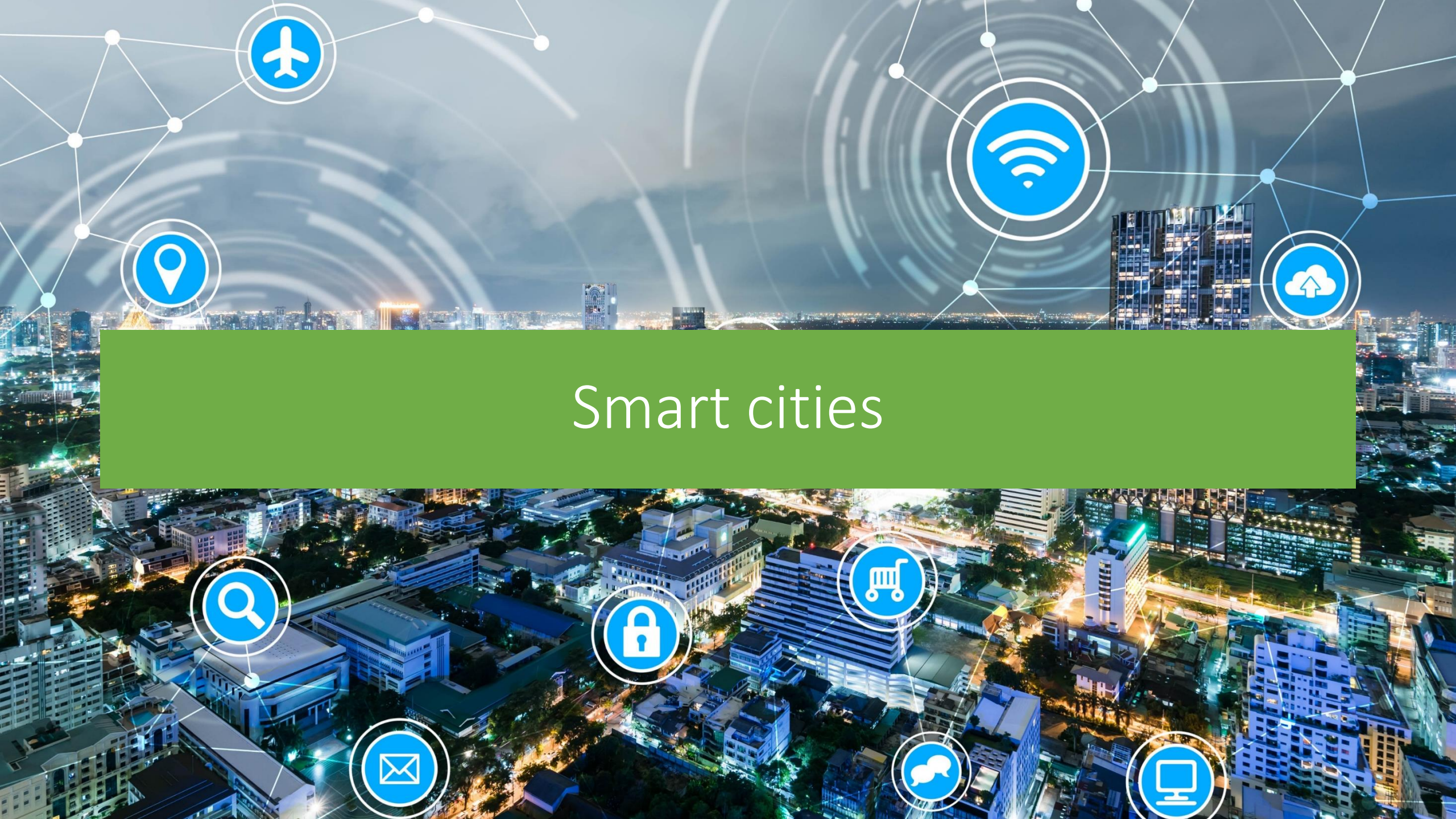


Automated decisions

Virtual and augmented reality

Domestic robots

Smart buildings



Smart cities

High-tech jobs

Problems

Unemployment

Simple, repetitive, and costly jobs
Unemployment

Simple, repetitive, and costly jobs
600 million jobs lost by 2030
Unemployment

Simple, repetitive, and costly jobs
600 million jobs lost by 2030
700 million new jobs by 2030
Unemployment

Simple, repetitive, and costly jobs

600 million jobs lost by 2030

700 million new jobs by 2030

Unemployment

Labor market is unprepared

Simple, repetitive, and costly jobs

600 million jobs lost by 2030

700 million new jobs by 2030

Unemployment

Labor market is unprepared

Society is unprepared

Simple, repetitive, and costly jobs

600 million jobs lost by 2030

700 million new jobs by 2030

Unemployment

Labor market is unprepared

Society is unprepared

Many people left behind

Synthetic media

Indistinguishable from humans

Synthetic media

Indistinguishable from humans

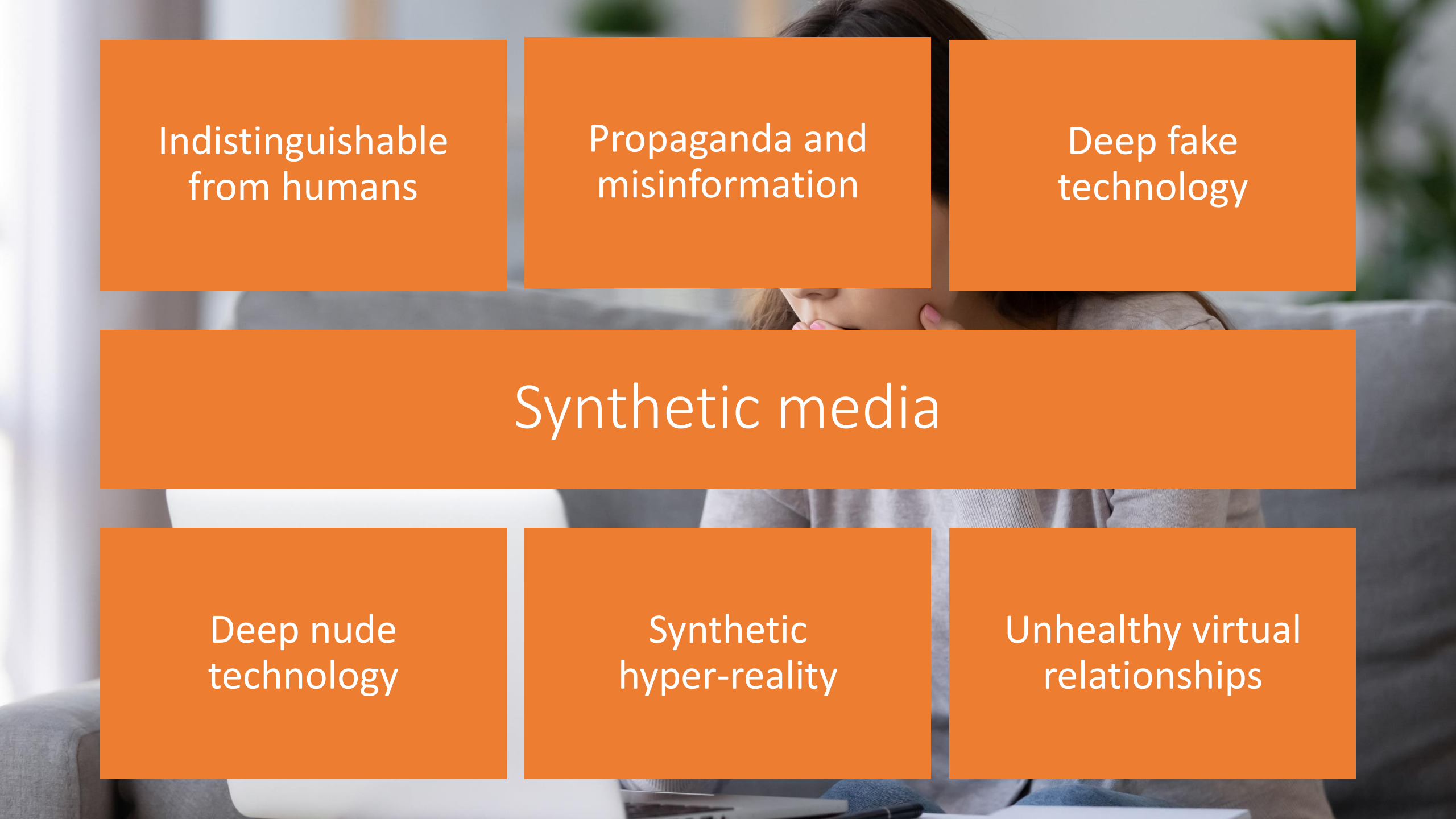
Propaganda and misinformation

Synthetic media

Indistinguishable from humans
Propaganda and misinformation
Deep fake technology
Synthetic media

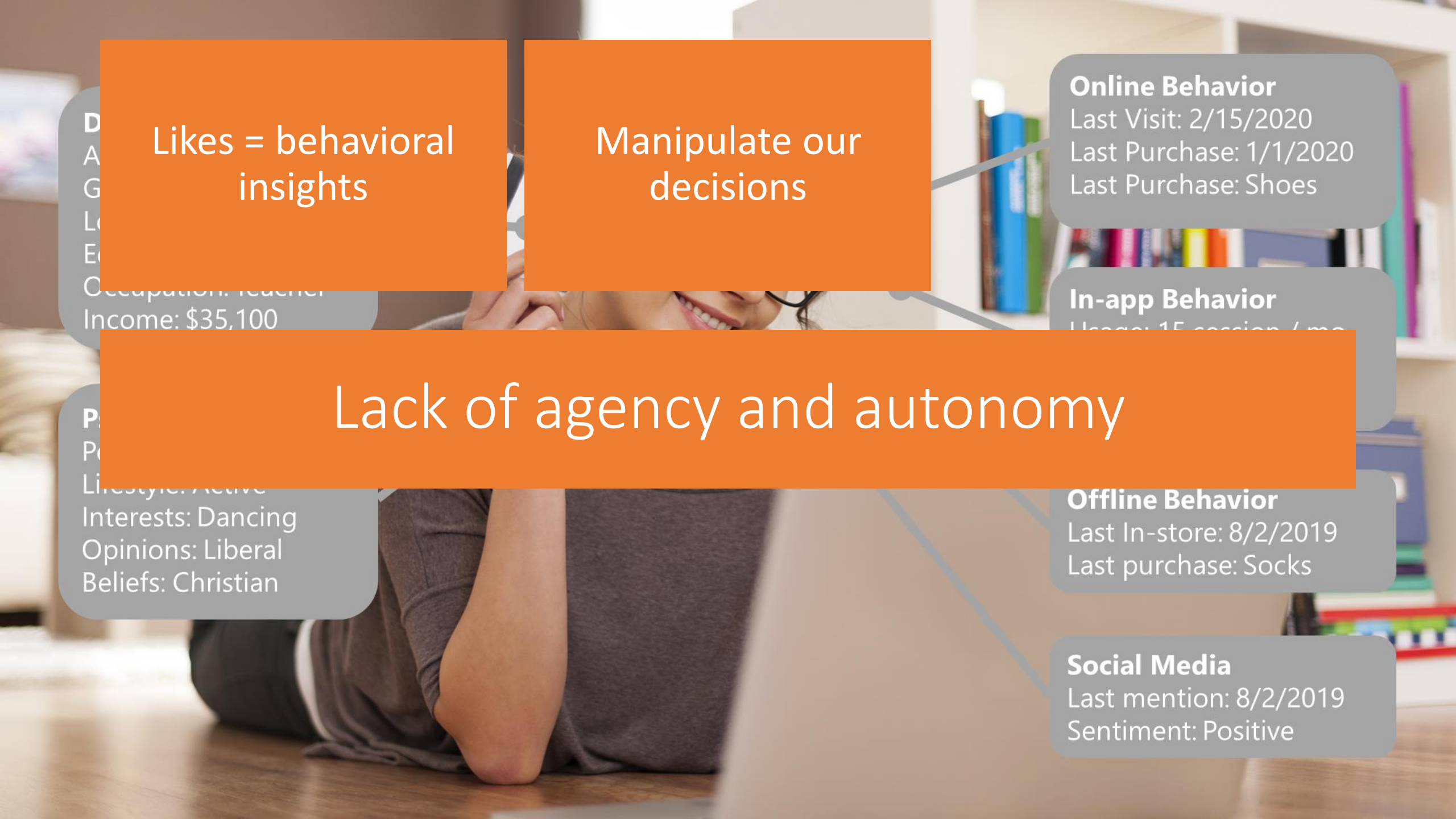
Indistinguishable from humans
Propaganda and misinformation
Deep fake technology
Synthetic media
Deep nude technology

Indistinguishable from humans
Propaganda and misinformation
Deep fake technology
Synthetic media
Deep nude technology
Synthetic hyper-reality

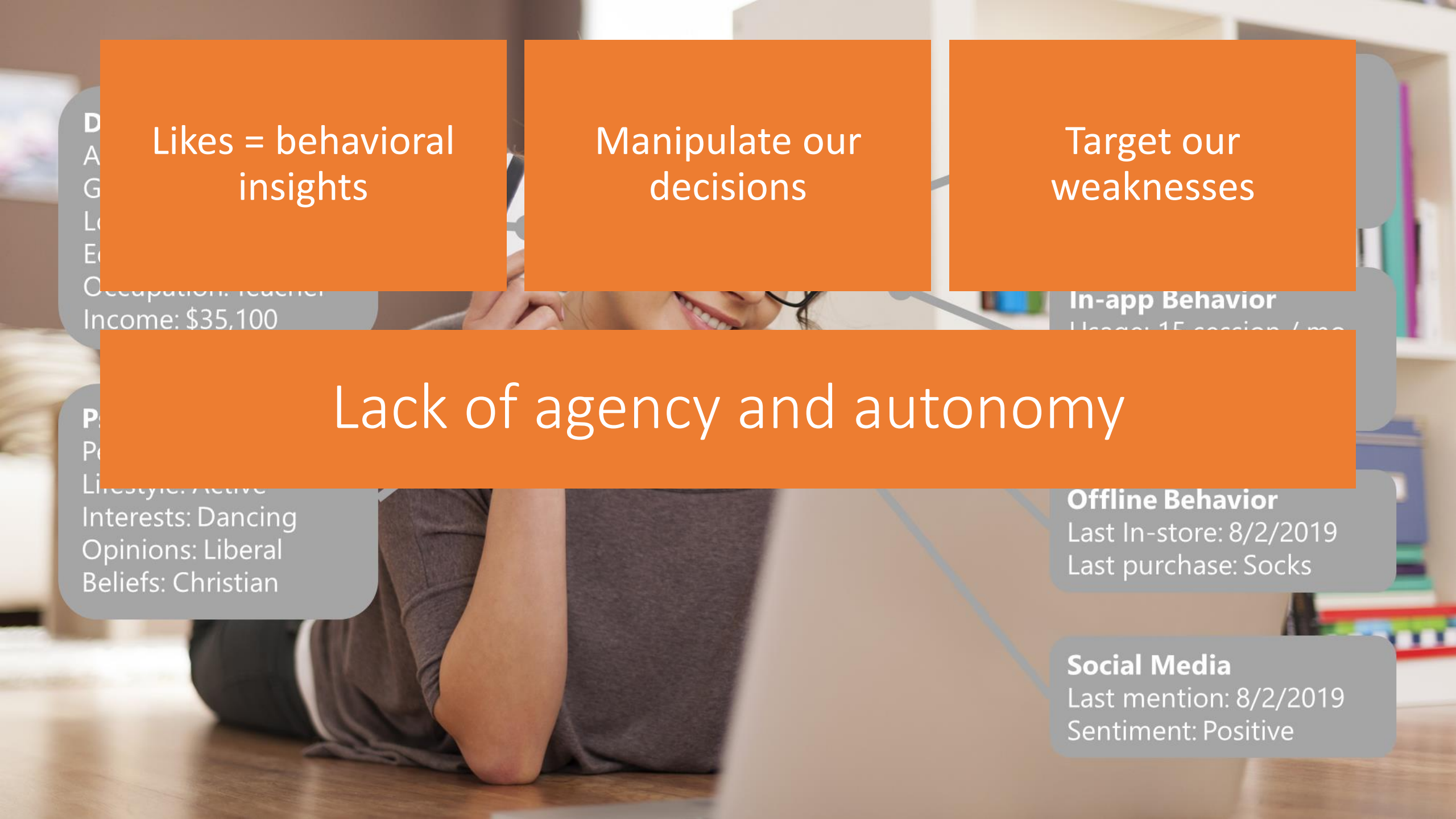

Indistinguishable from humans
Propaganda and misinformation
Deep fake technology
Synthetic media
Deep nude technology
Synthetic hyper-reality
Unhealthy virtual relationships

Demographics
Age: 20
Gender: Female
Location: Chicago
Education: MS in Edu.
Occupation: Teacher
Income: $35,100
Online Behavior
Last Visit: 2/15/2020
Last Purchase: 1/1/2020
Last Purchase: Shoes
In-app Behavior
Interests: Dancing
Opinions: Liberal
Beliefs: Christian
Offline Behavior
Last In-store: 8/2/2019
Last purchase: Socks
Social Media
Last mention: 8/2/2019
Sentiment: Positive
Lack of agency and autonomy

Lack of agency and autonomy



Lack of agency and autonomy



Lack of agency and autonomy

Likes = behavioral insights
Manipulate our decisions
Target our weaknesses
Income: $35,100
In-app Behavior
Lifestyle: Active
Interests: Dancing
Offline Behavior
Last In-store: 8/2/2019
Last purchase: Socks
Social Media
Last mention: 8/2/2019
Sentiment: Positive
Lack of agency and autonomy
Surveillance capitalism

Likes = behavioral insights
Manipulate our decisions
Target our weaknesses
Income: $35,100
In-app Behavior
Lack of agency and autonomy
Interests: Dancing
Offline Behavior
Last In-store: 8/2/2019
Last purchase: Socks
Surveillance capitalism
Lack of freedom
Social Media
Last mention: 8/2/2019
Sentiment: Positive

Likes = behavioral insights
Manipulate our decisions
Target our weaknesses
Income: $35,100
In-app Behavior
Interests: Dancing
Offline Behavior
Lack of agency and autonomy
Surveillance capitalism
Lack of freedom
Forced compliance

Economic impacts

Social stratification

Safety issues

Autonomous weapons

Solutions

Educate everyone

What is AI?

Educate everyone

What is AI?

Basics of
ML, DL, RL

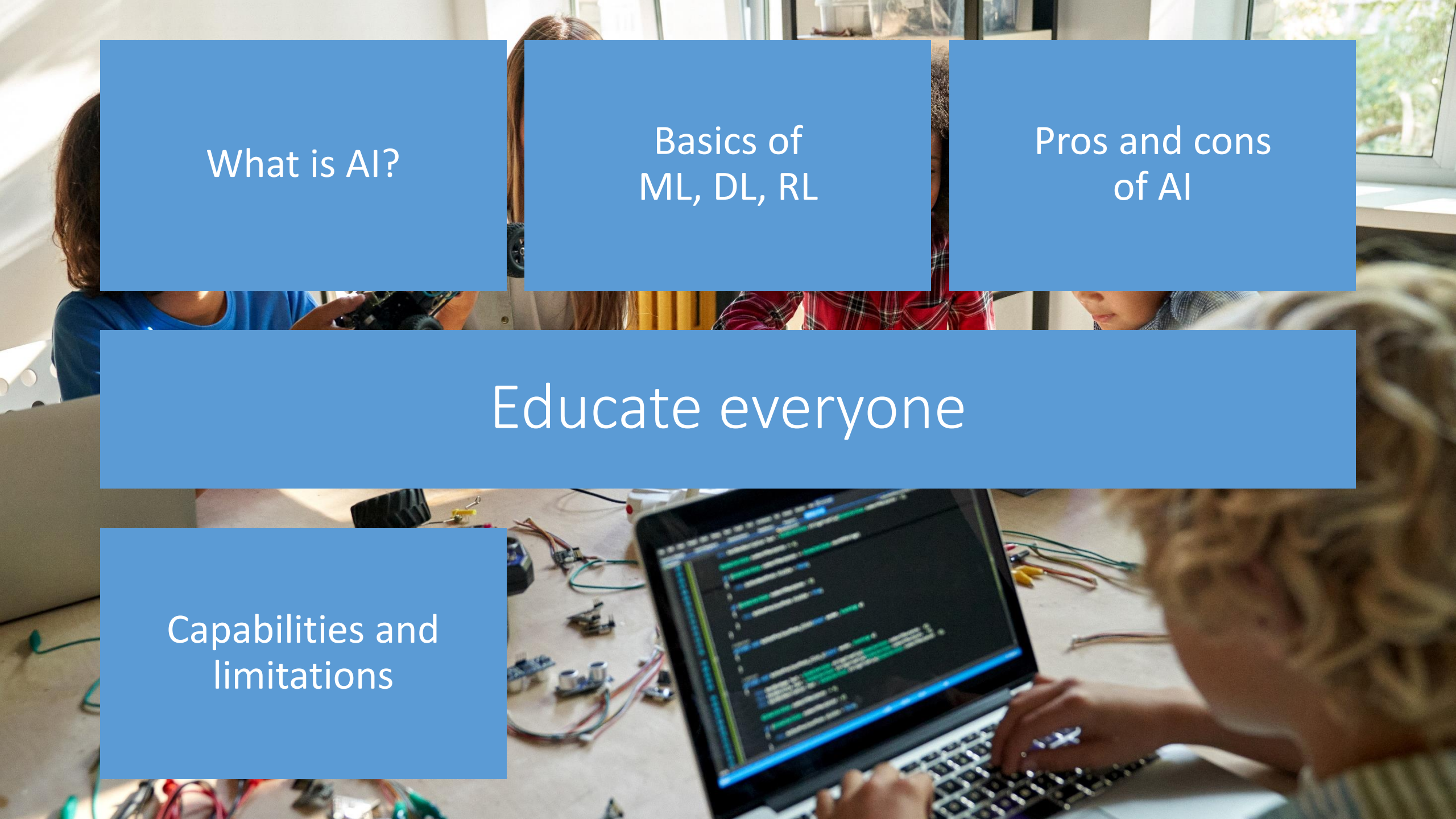
Educate everyone

What is AI?

Basics of
ML, DL, RL

Pros and cons
of AI

Educate everyone



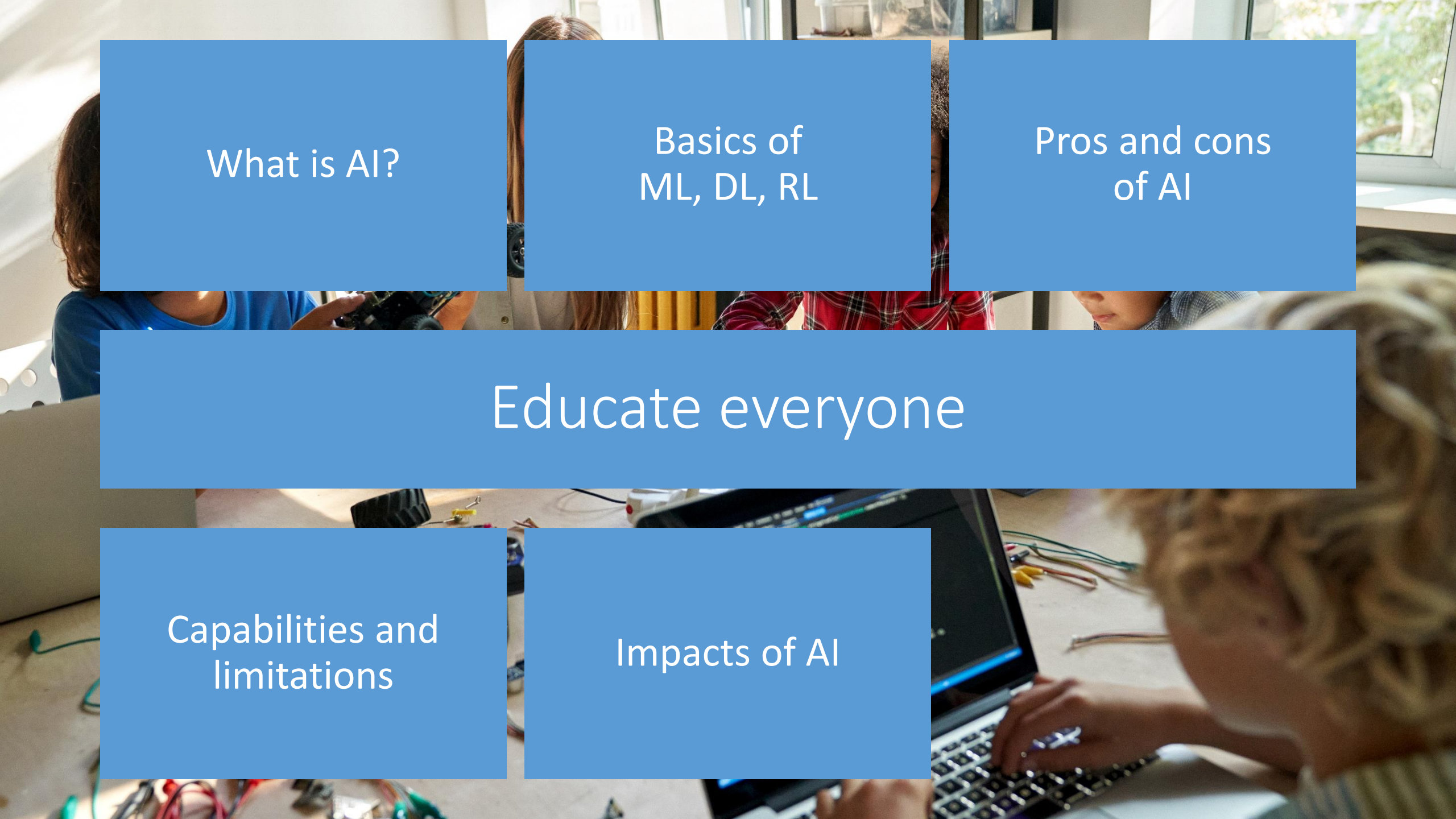
What is AI?

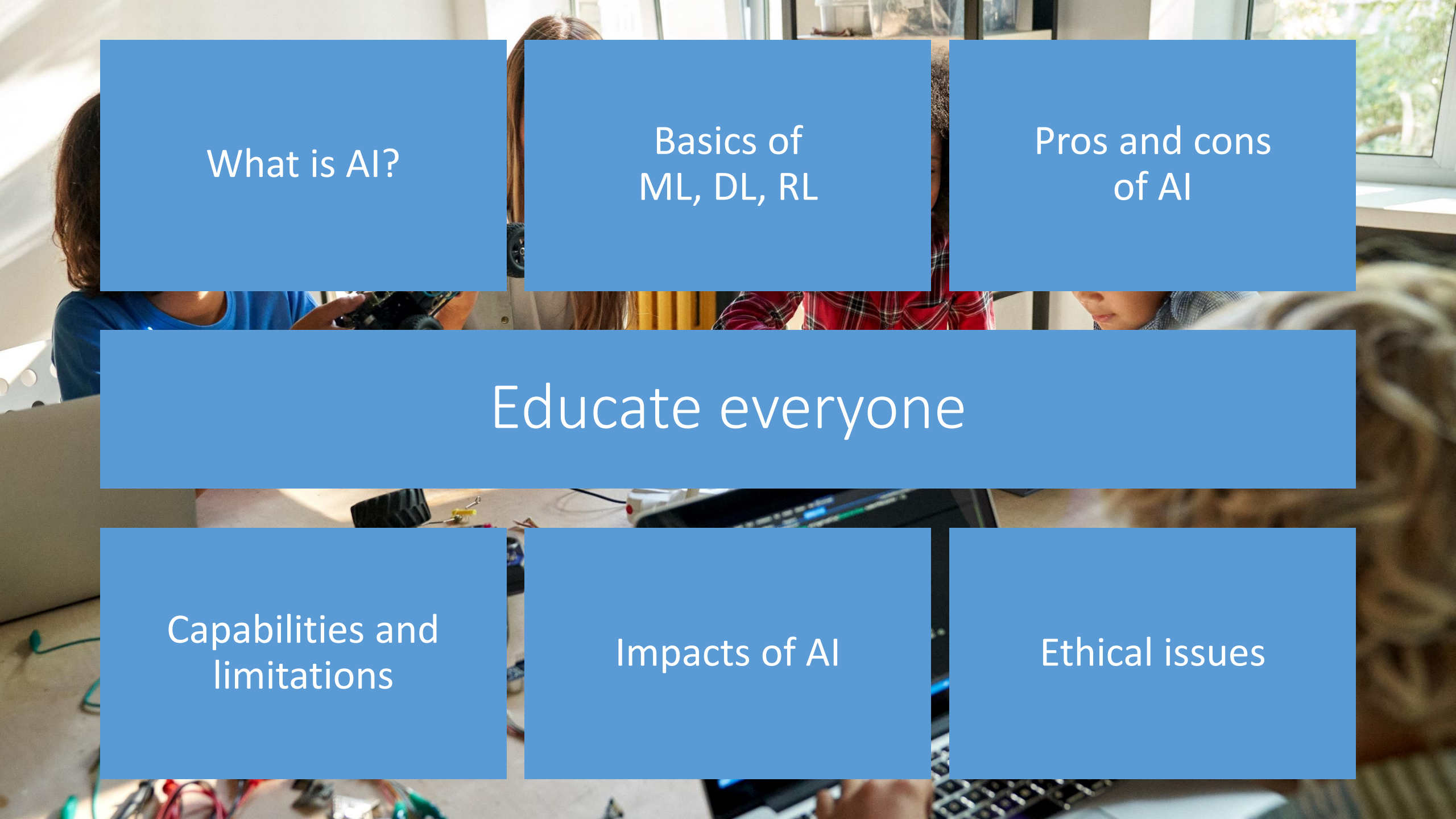
Basics of
ML, DL, RL

Pros and cons
of AI

Educate everyone

Capabilities and
limitations


What is AI?
Basics of
ML, DL, RL
Pros and cons
of AI
Educate everyone
Capabilities and
limitations
Impacts of AI


What is AI?
Basics of ML, DL, RL
Pros and cons of AI
Educate everyone
Capabilities and limitations
Impacts of AI
Ethical issues

Upgrade skills

Job retraining programs

Upgrade skills

Job retraining programs

Retirement options

Upgrade skills

Job retraining programs

Retirement options

Social safety nets

Upgrade skills

Job retraining programs

Retirement options

Social safety nets

Upgrade skills

Mandatory higher education

Job retraining programs

Retirement options

Social safety nets

Upgrade skills

Mandatory higher education

Redistribute wealth and resources

Job retraining programs

Retirement options

Social safety nets

Upgrade skills

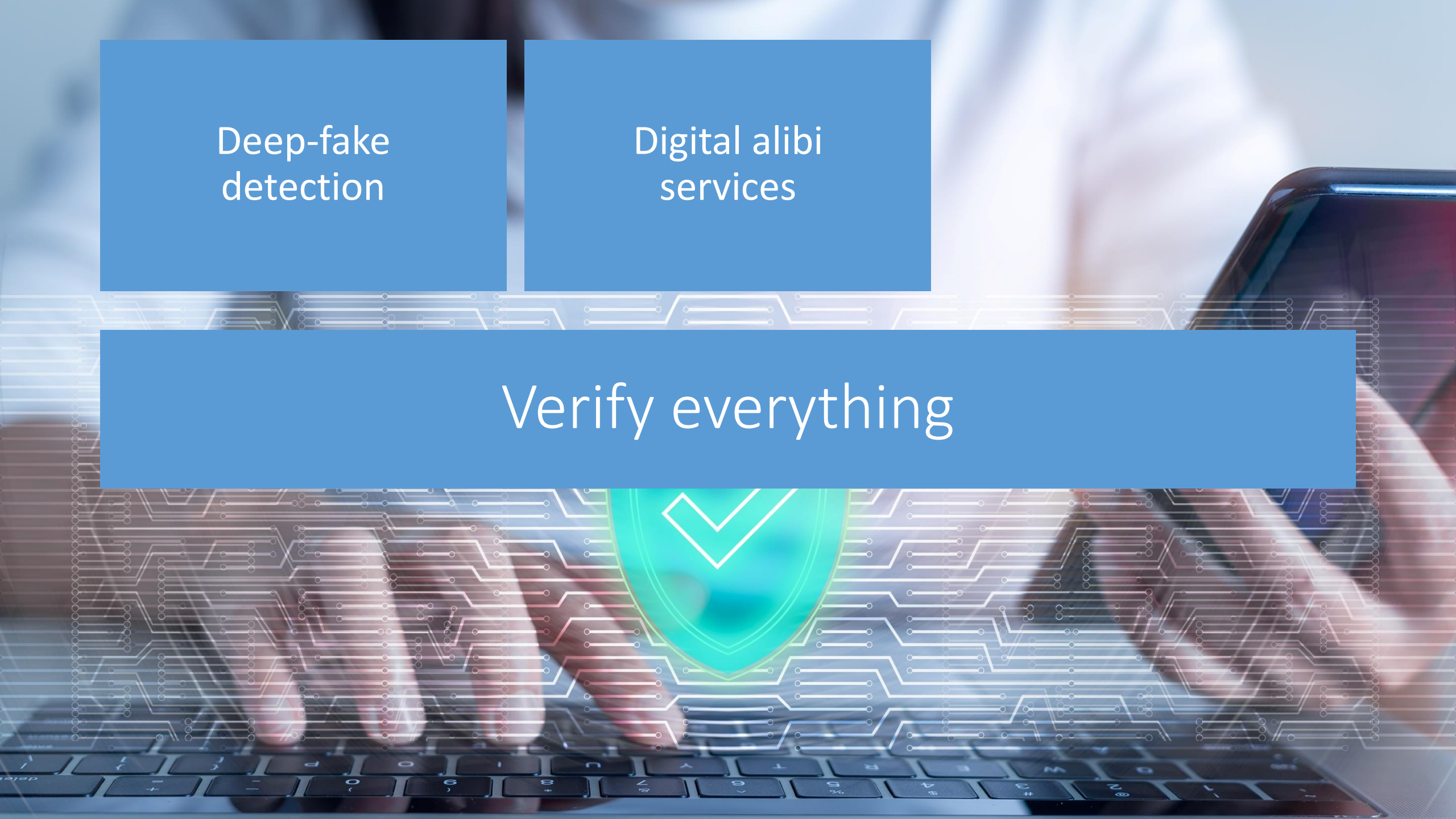
Mandatory higher education

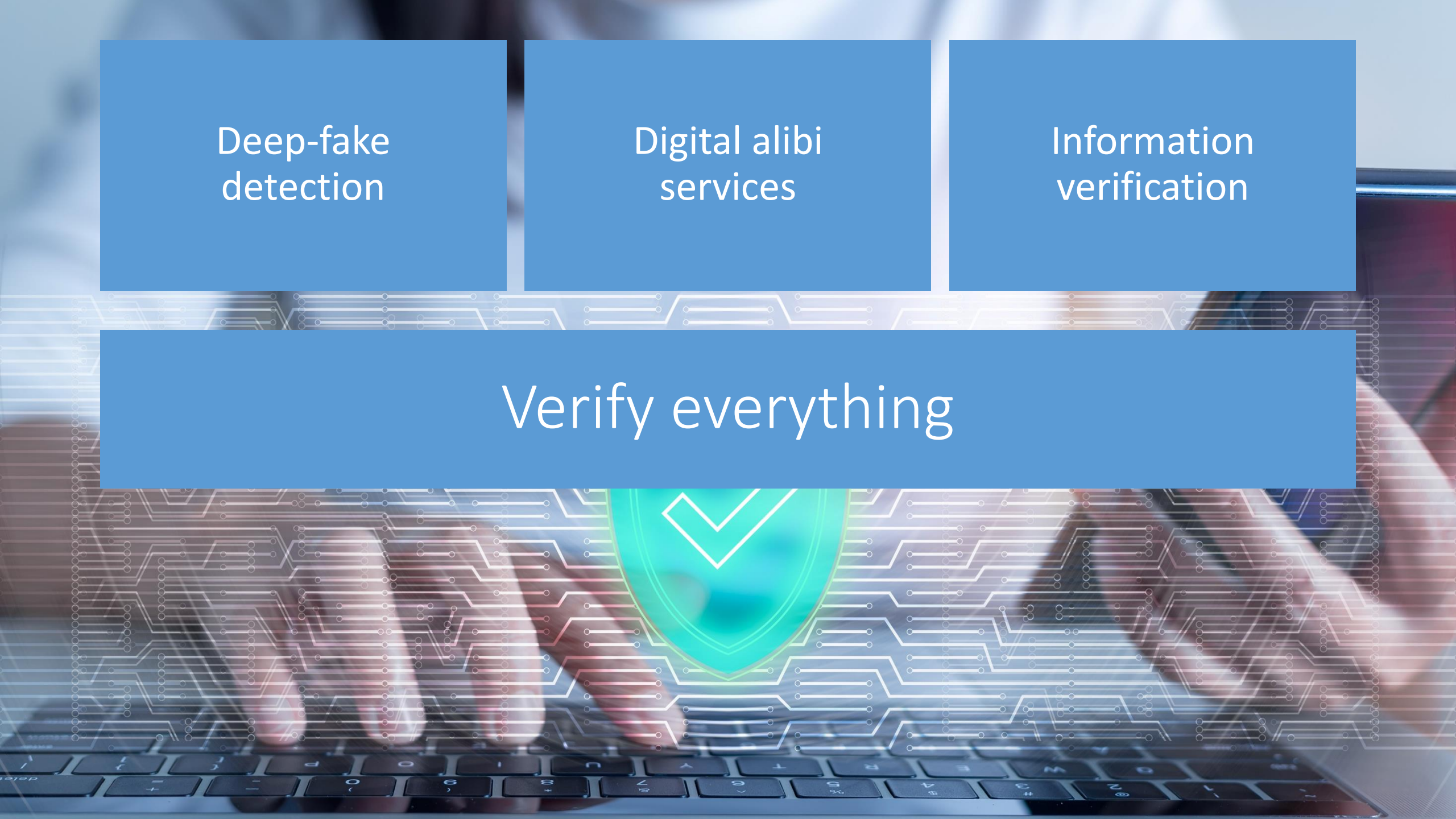
Redistribute wealth and resources

Universal basic income

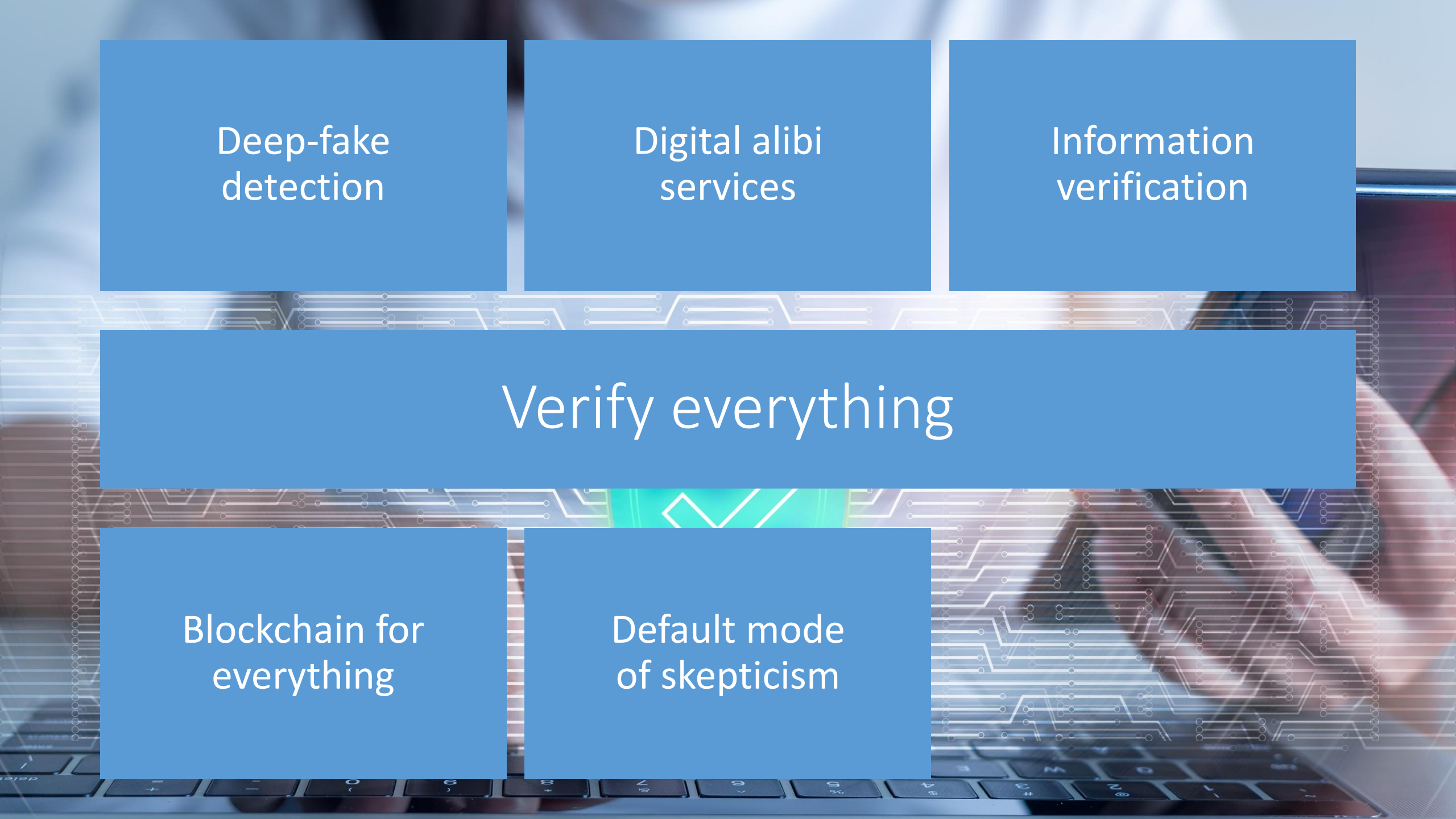
Verify everything

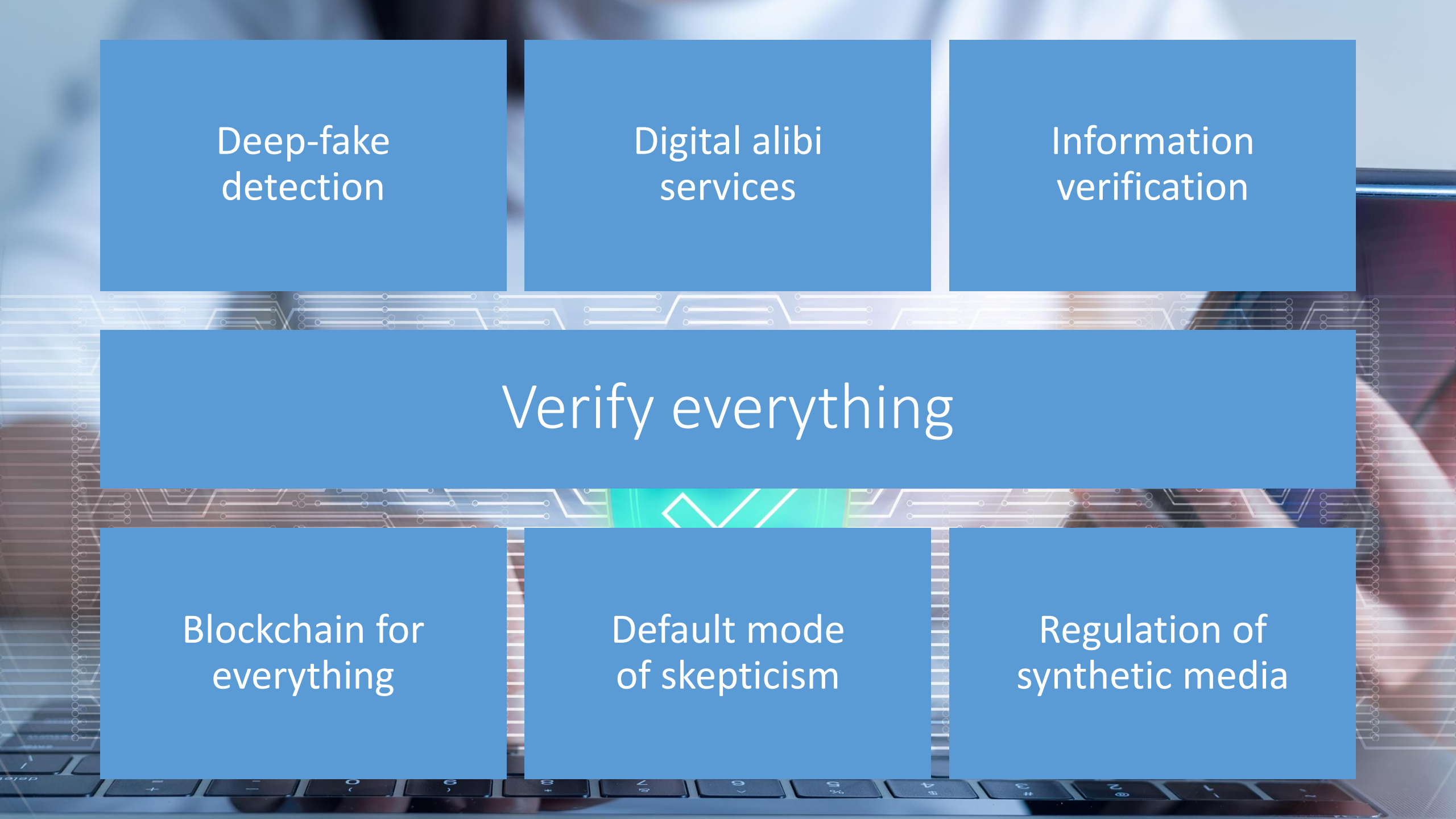
Deep-fake detection
Verify everything


Deep-fake detection
Digital alibi services
Verify everything


Deep-fake detection
Digital alibi services
Information verification
Verify everything

Deep-fake detection
Digital alibi services
Information verification
Verify everything
Blockchain for everything


Deep-fake detection
Digital alibi services
Information verification
Verify everything
Blockchain for everything
Default mode of skepticism


Deep-fake detection
Digital alibi services
Information verification
Verify everything
Blockchain for everything
Default mode of skepticism
Regulation of synthetic media

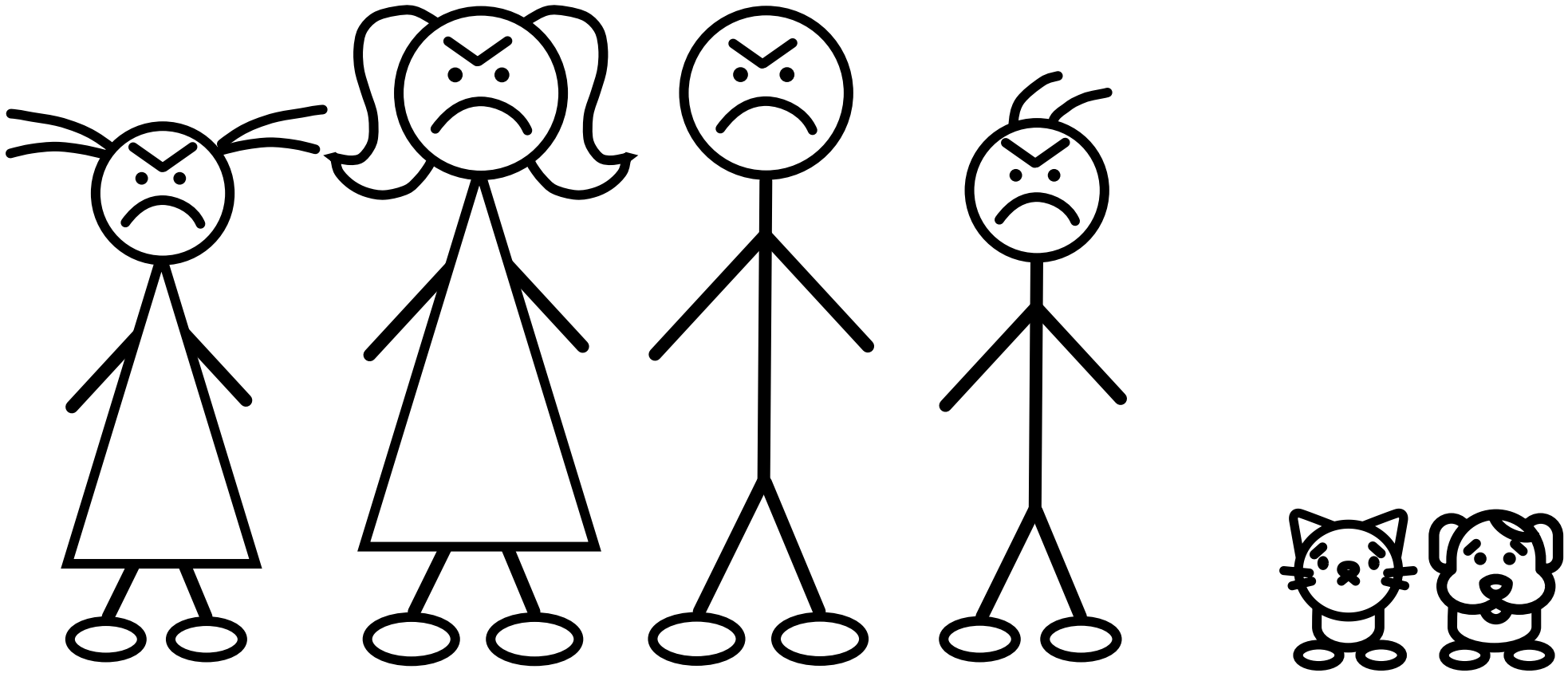
Maintain our agency and autonomy

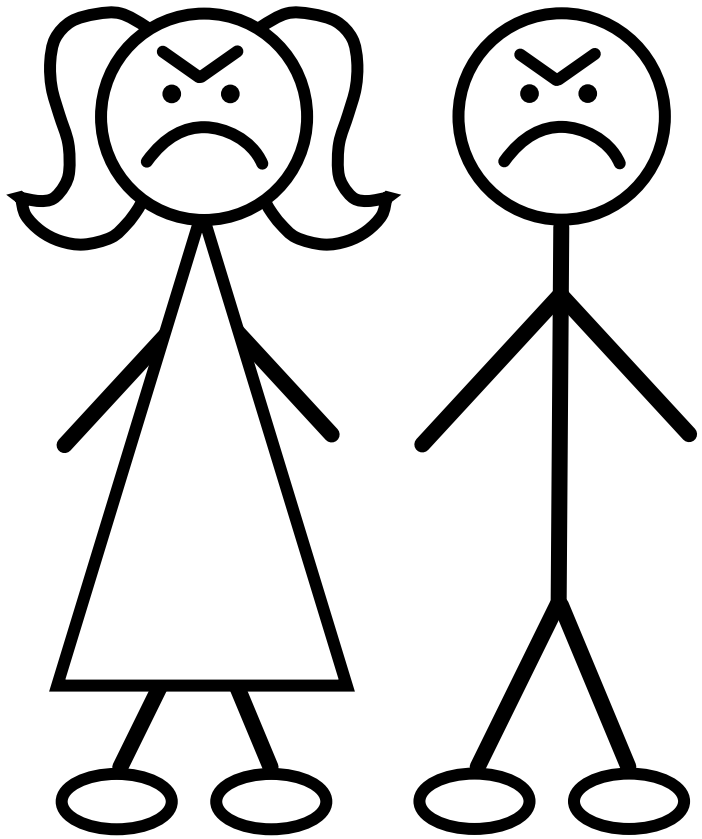
Implement safety solutions

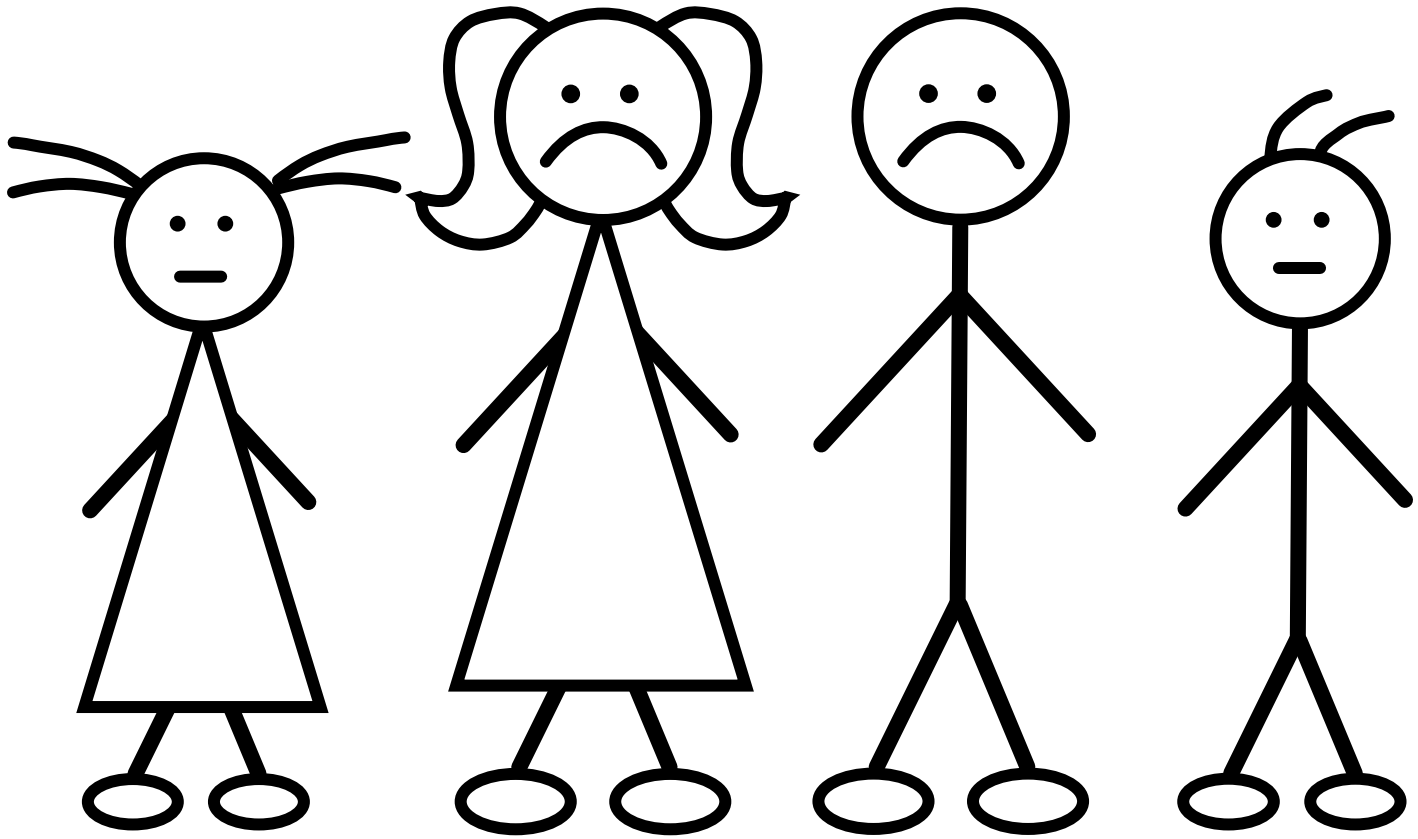
Ban autonomous weapons

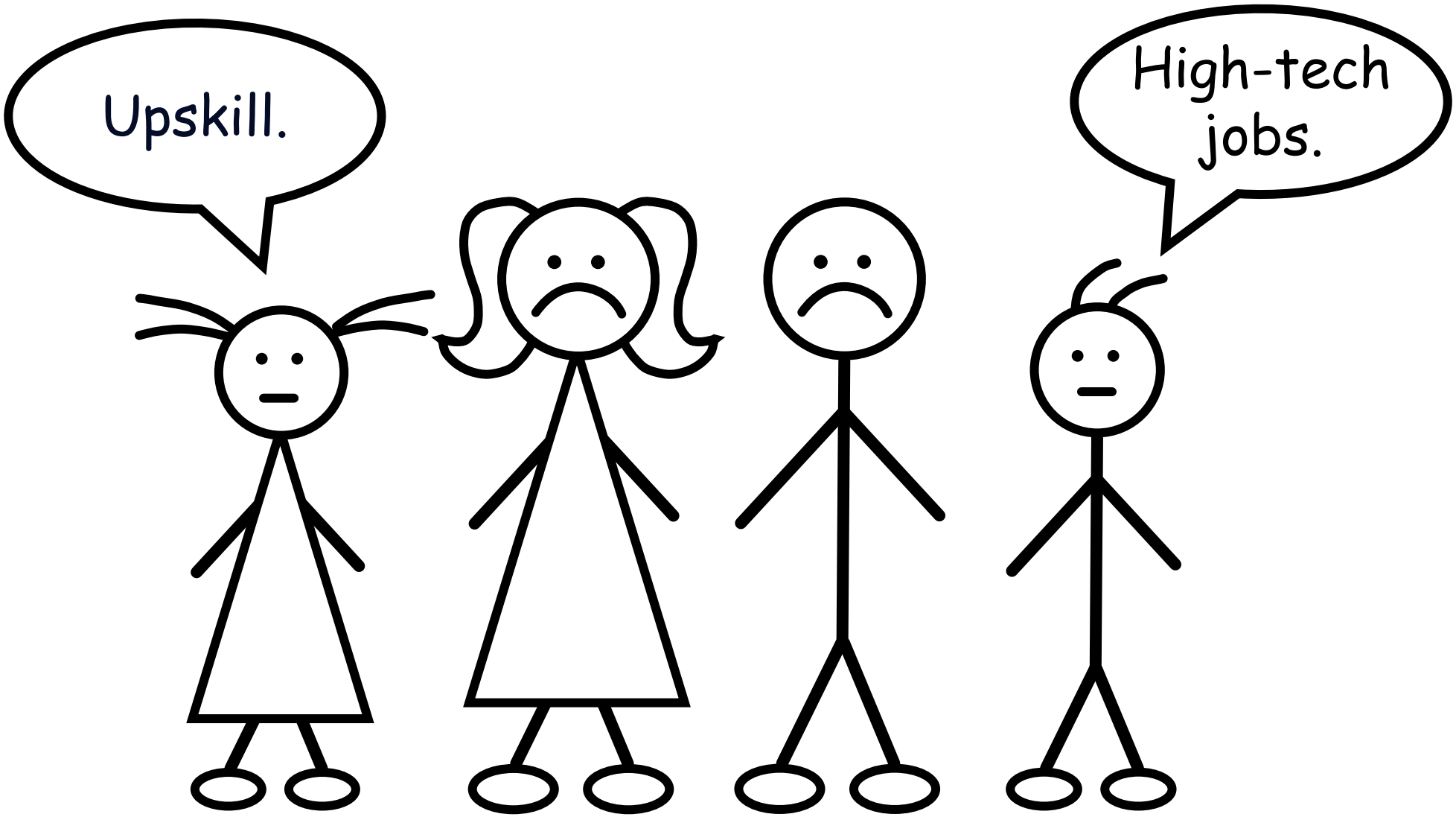
Alice and Bob in AI Land

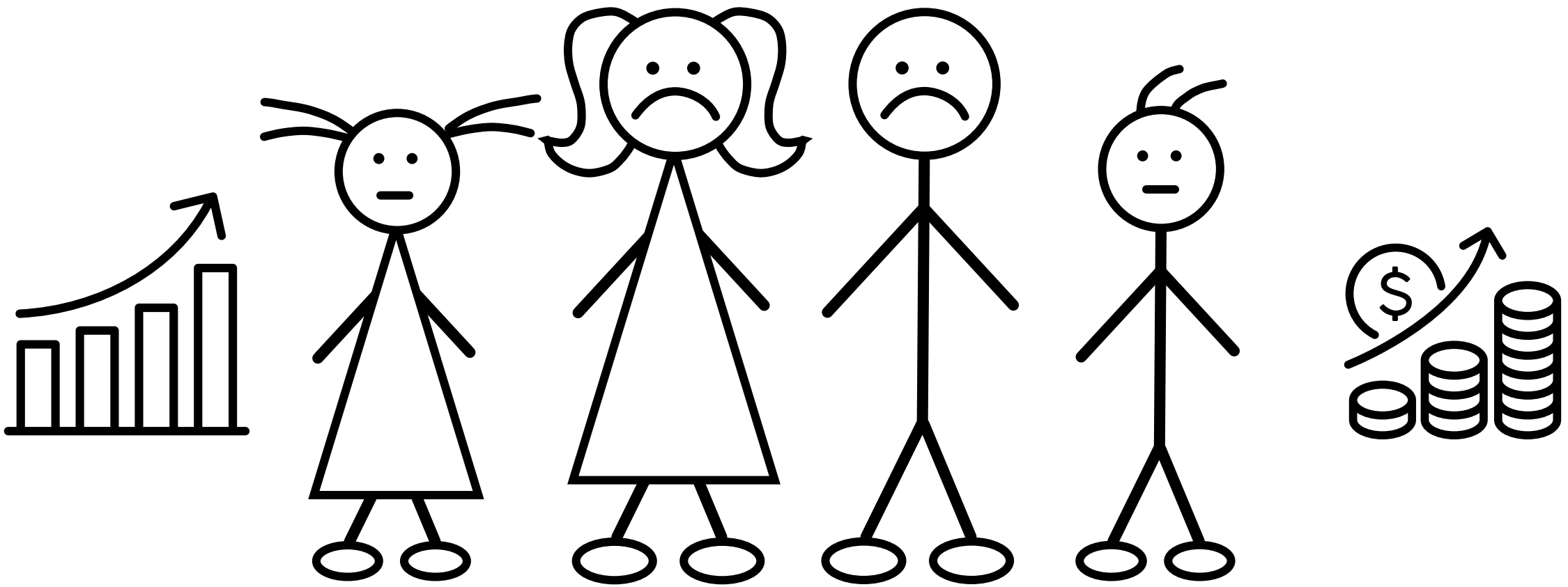
(continued …)

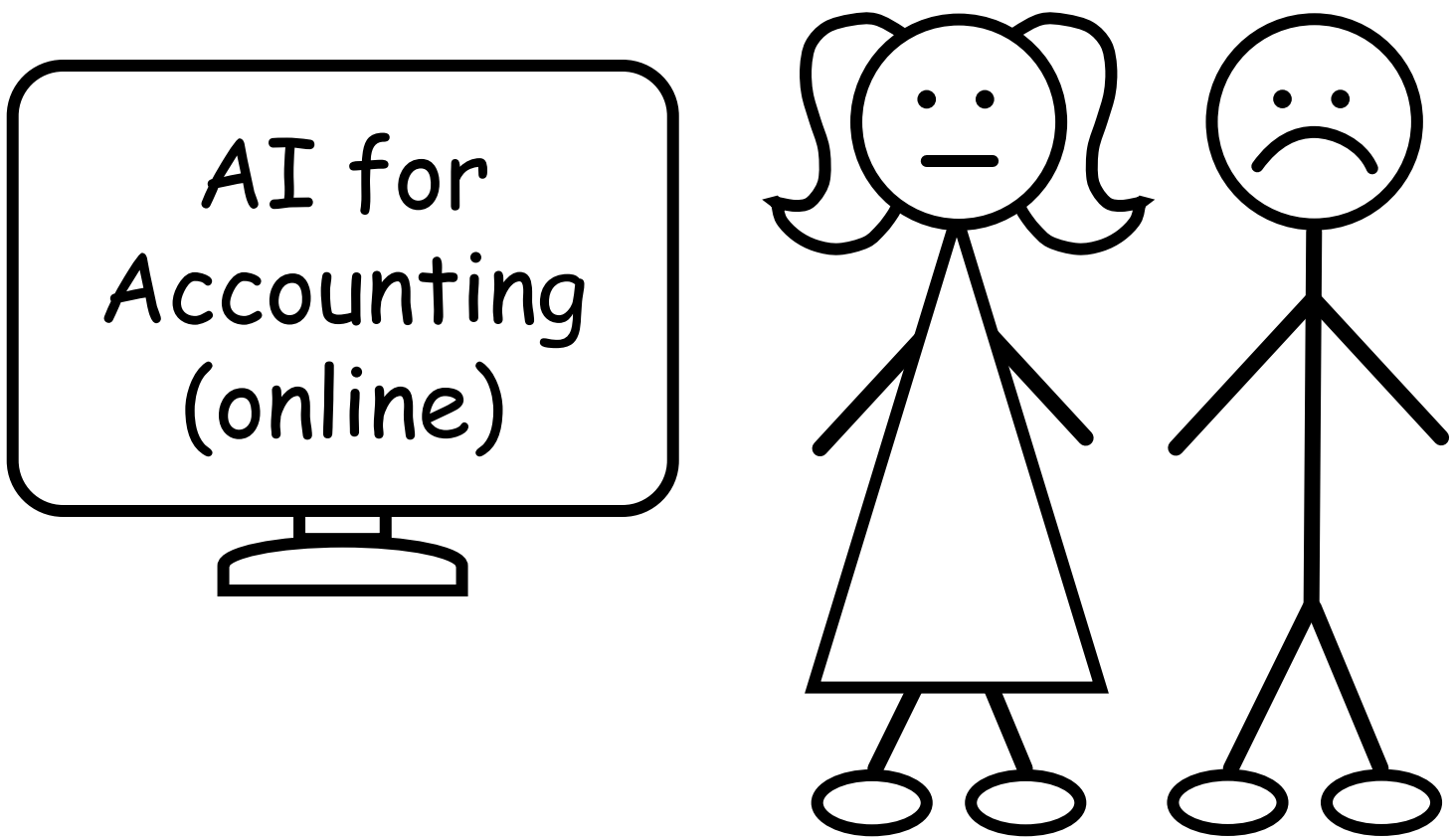


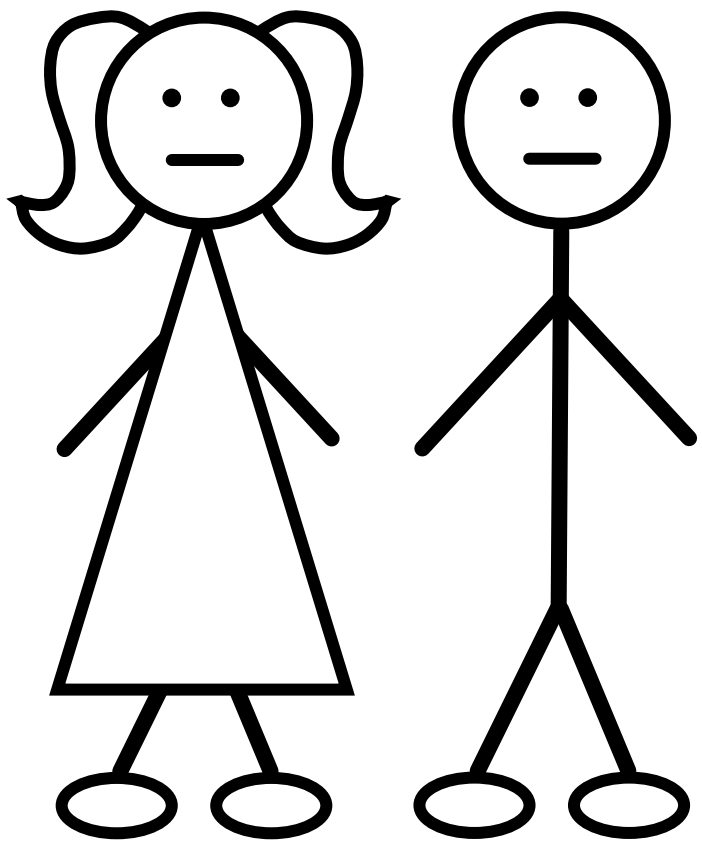




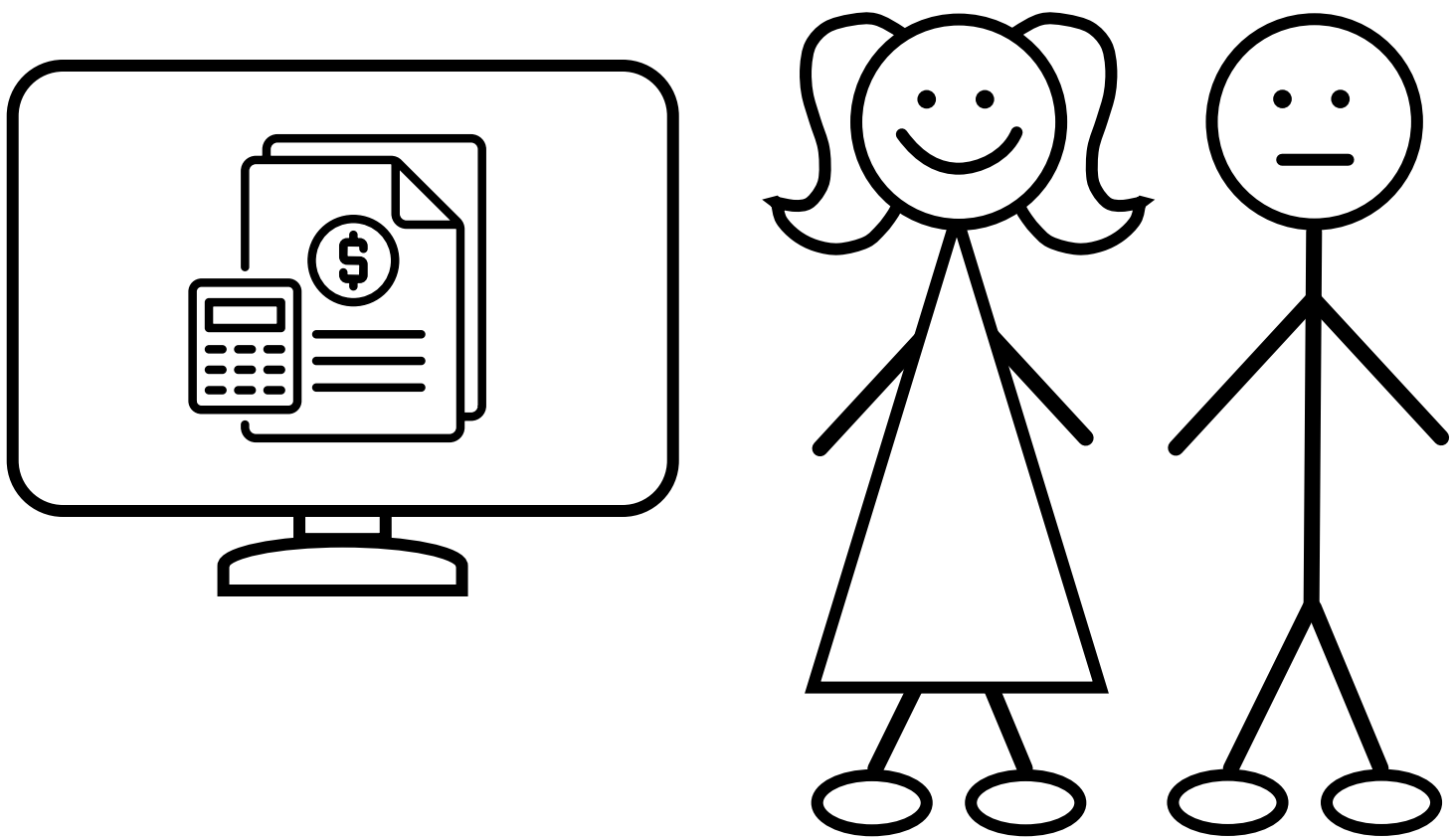

Upskill.
High-tech jobs.

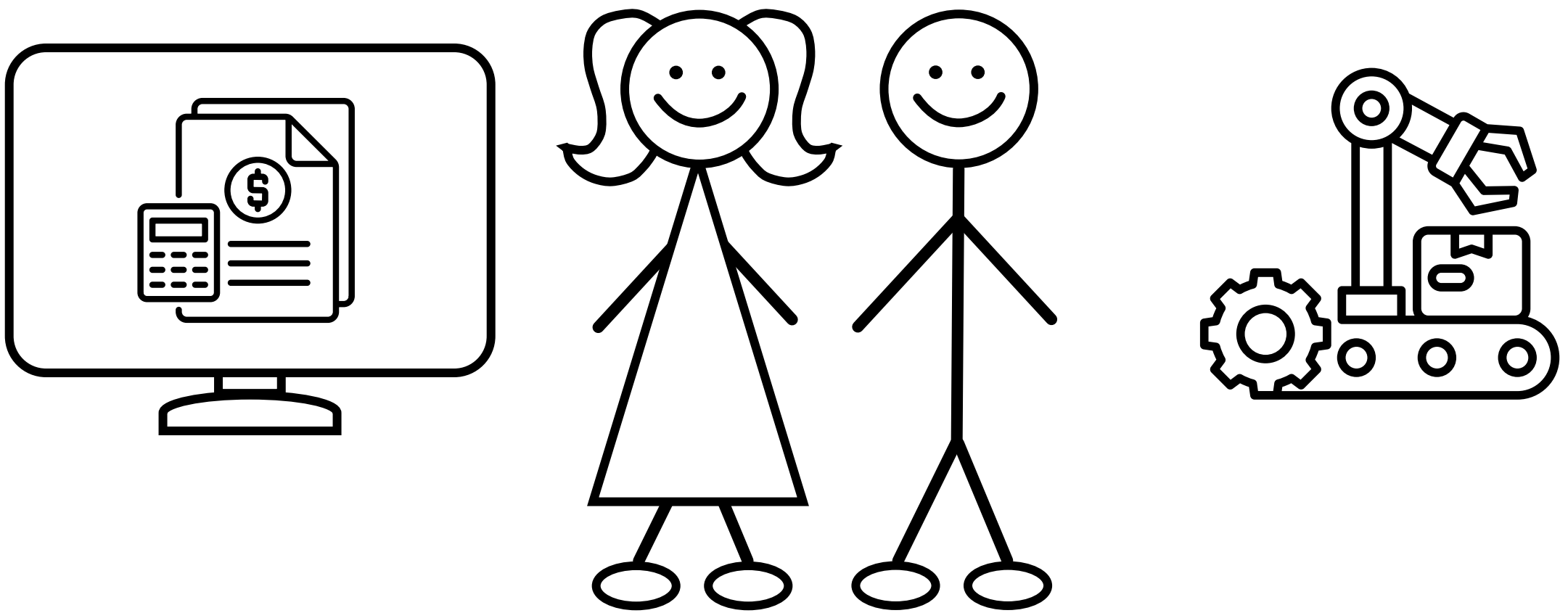


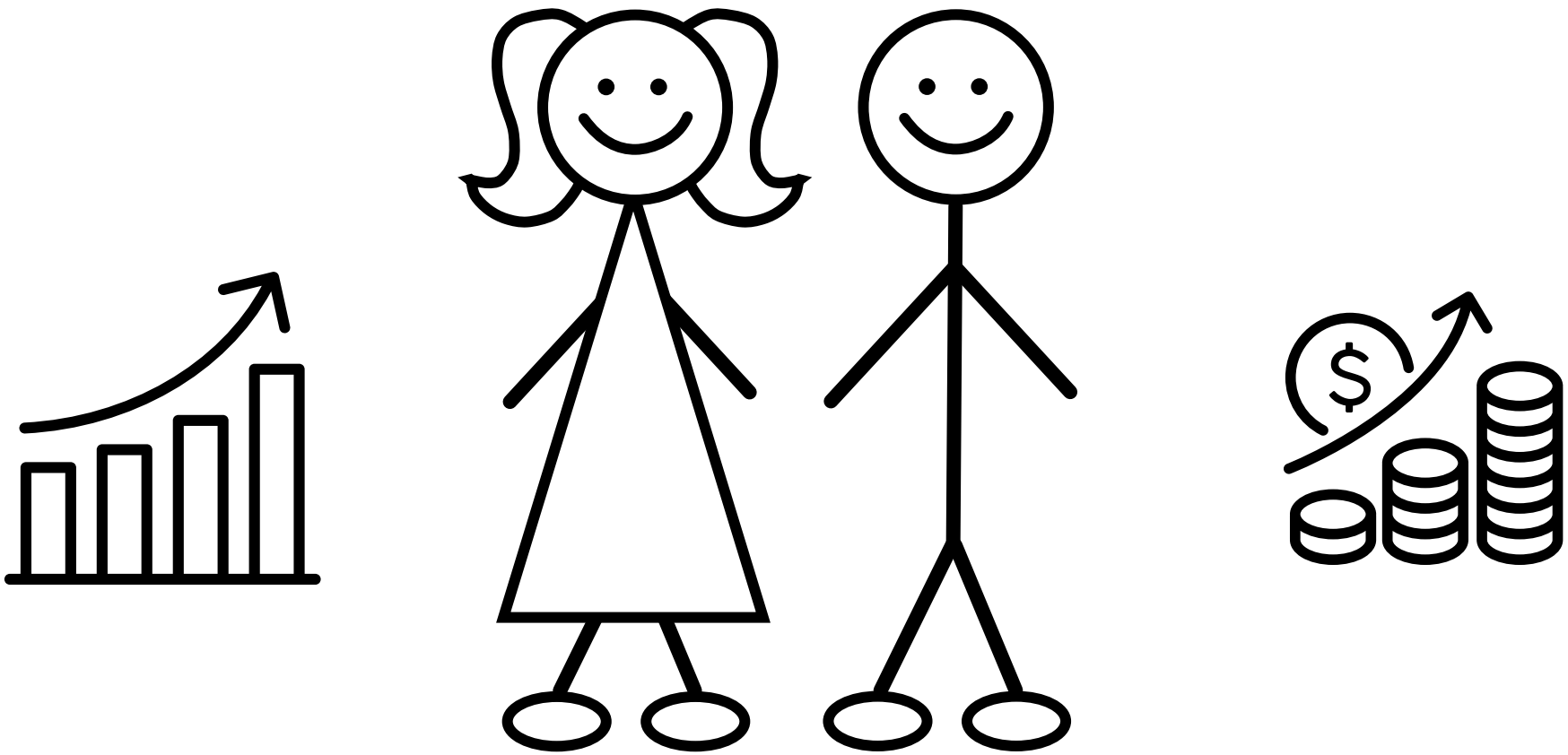

AI for
Accounting
(online)

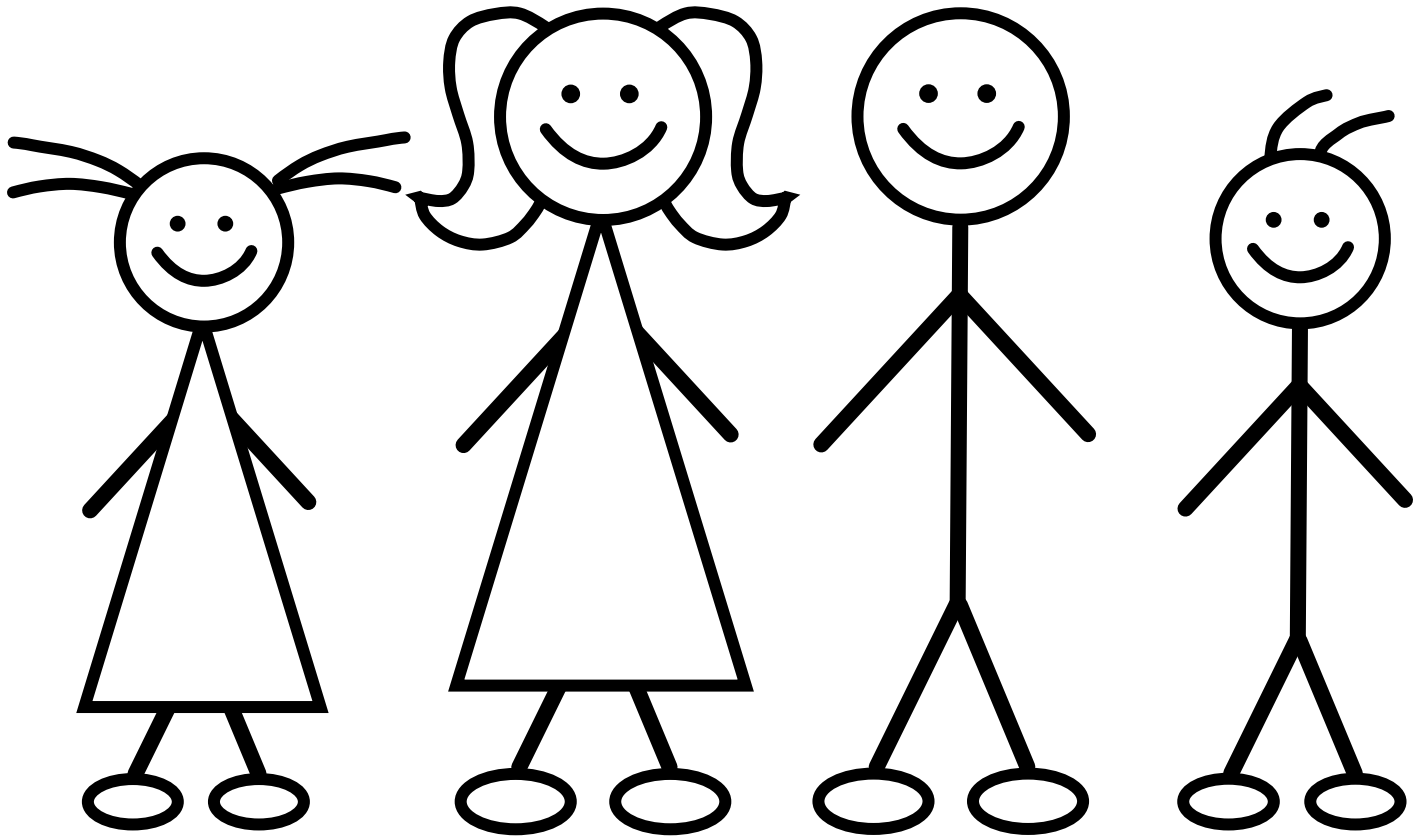


Robotics
Technician

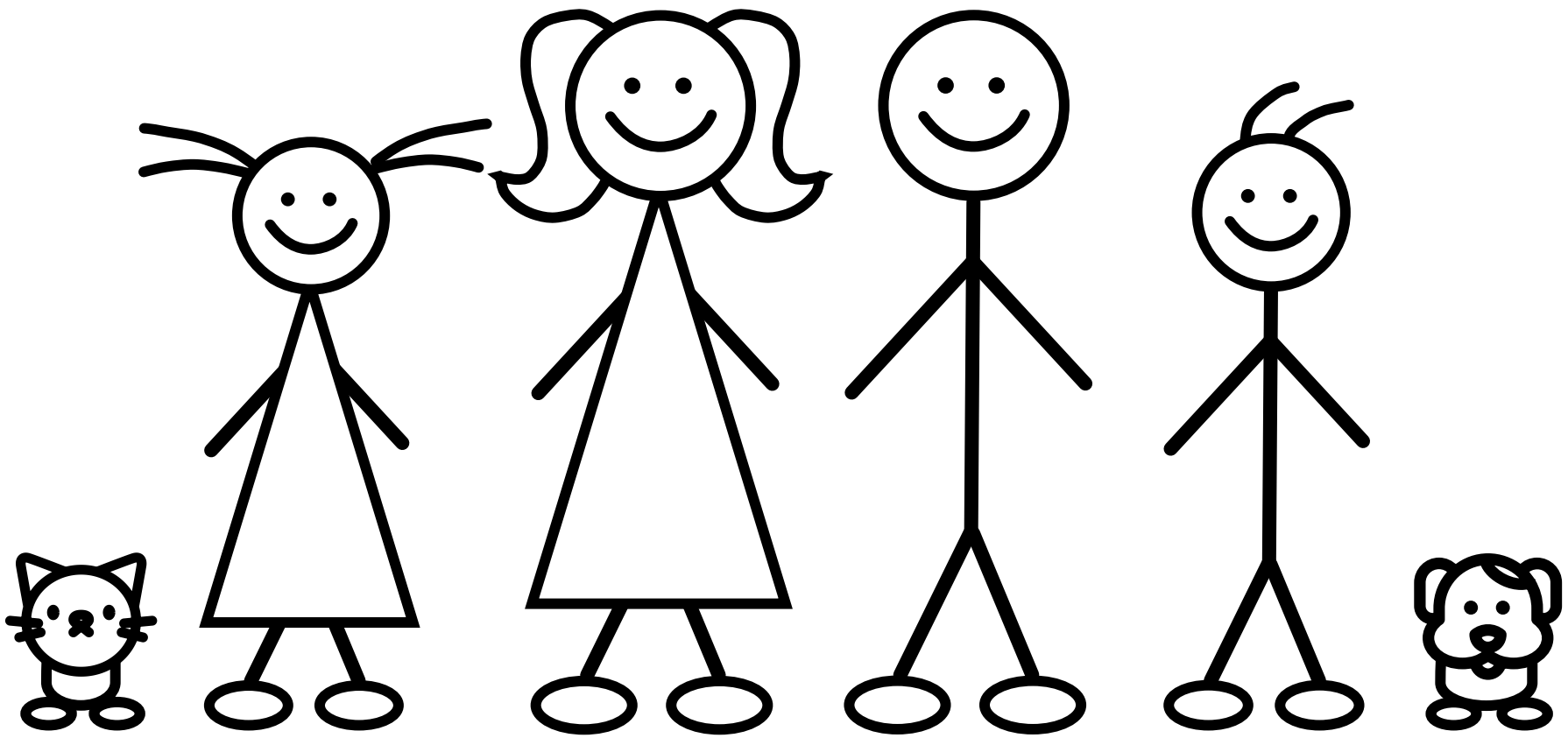








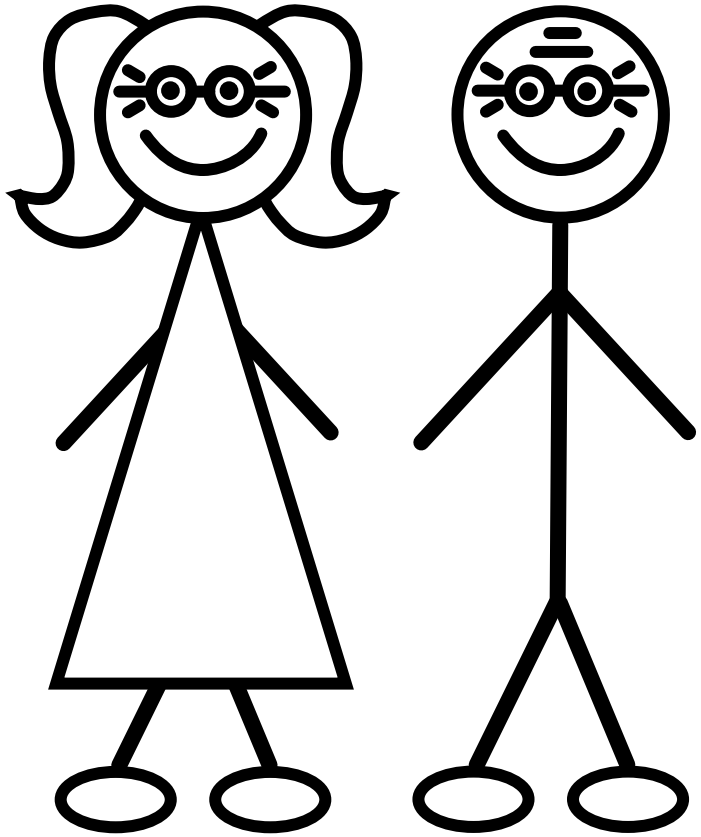
Happily ever after…

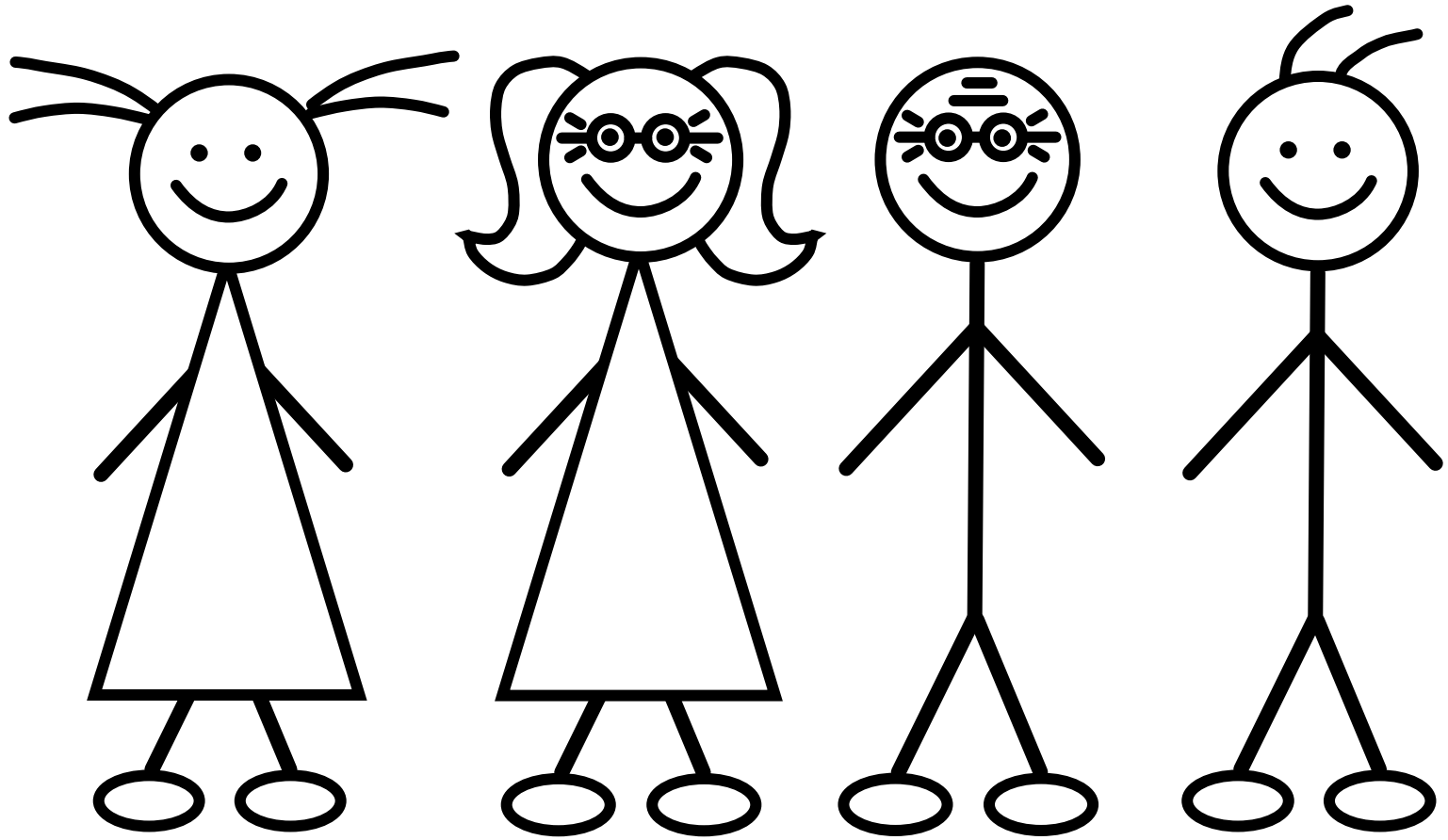


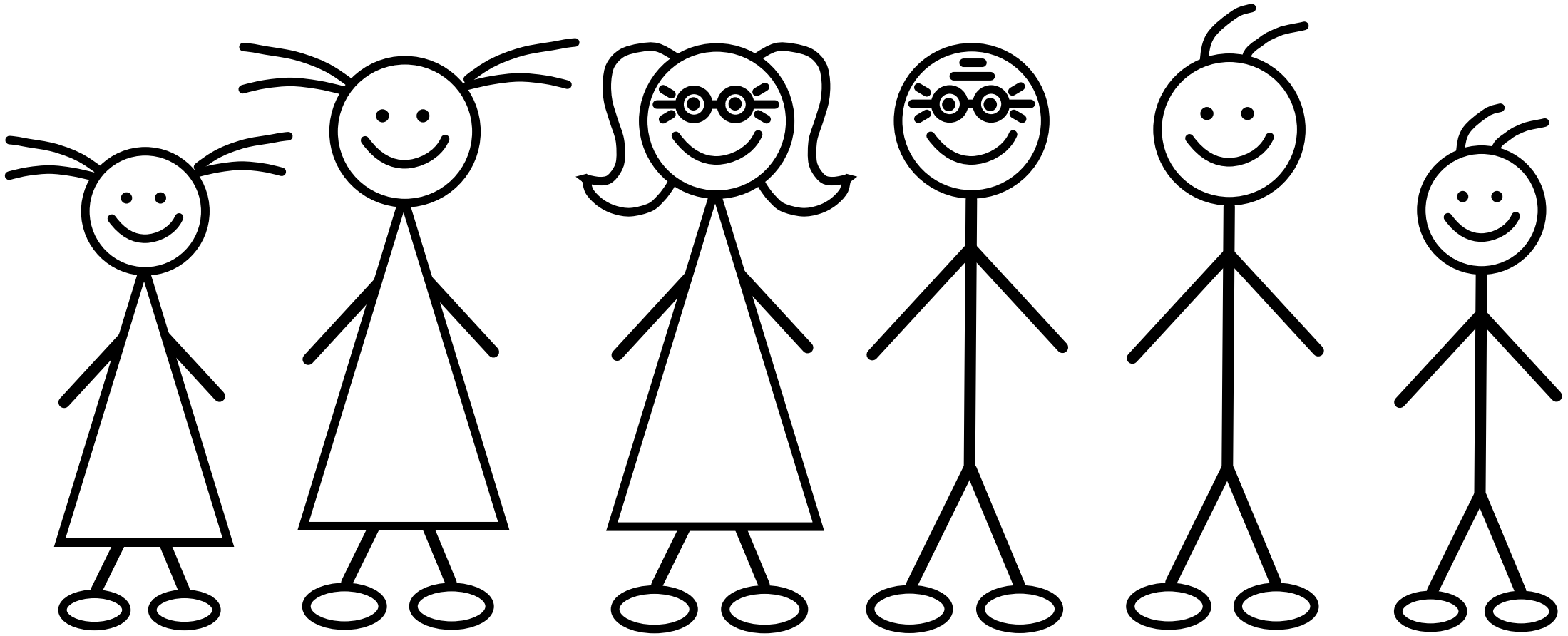
AI Ethics for The Future

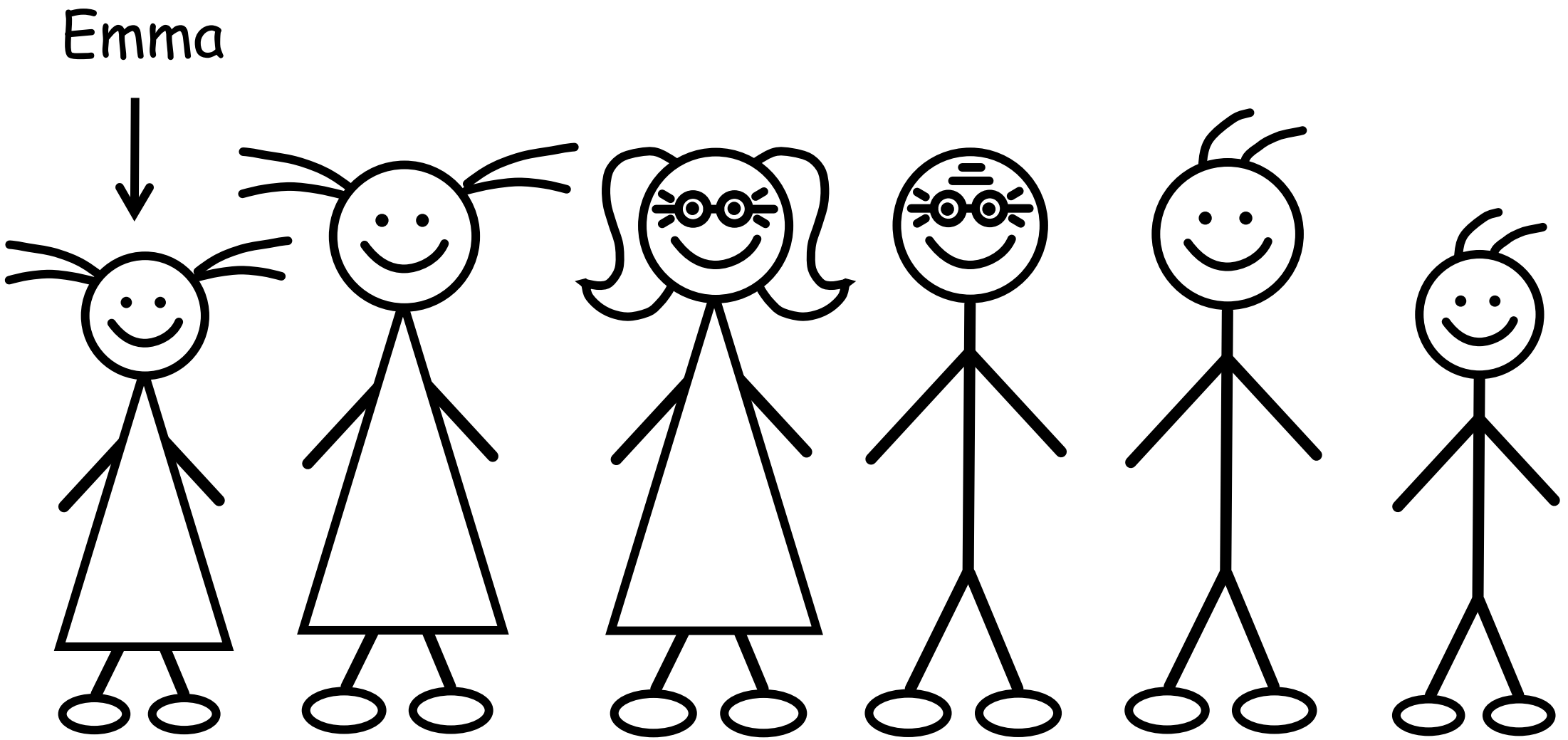
Alice and Bob in AI Land

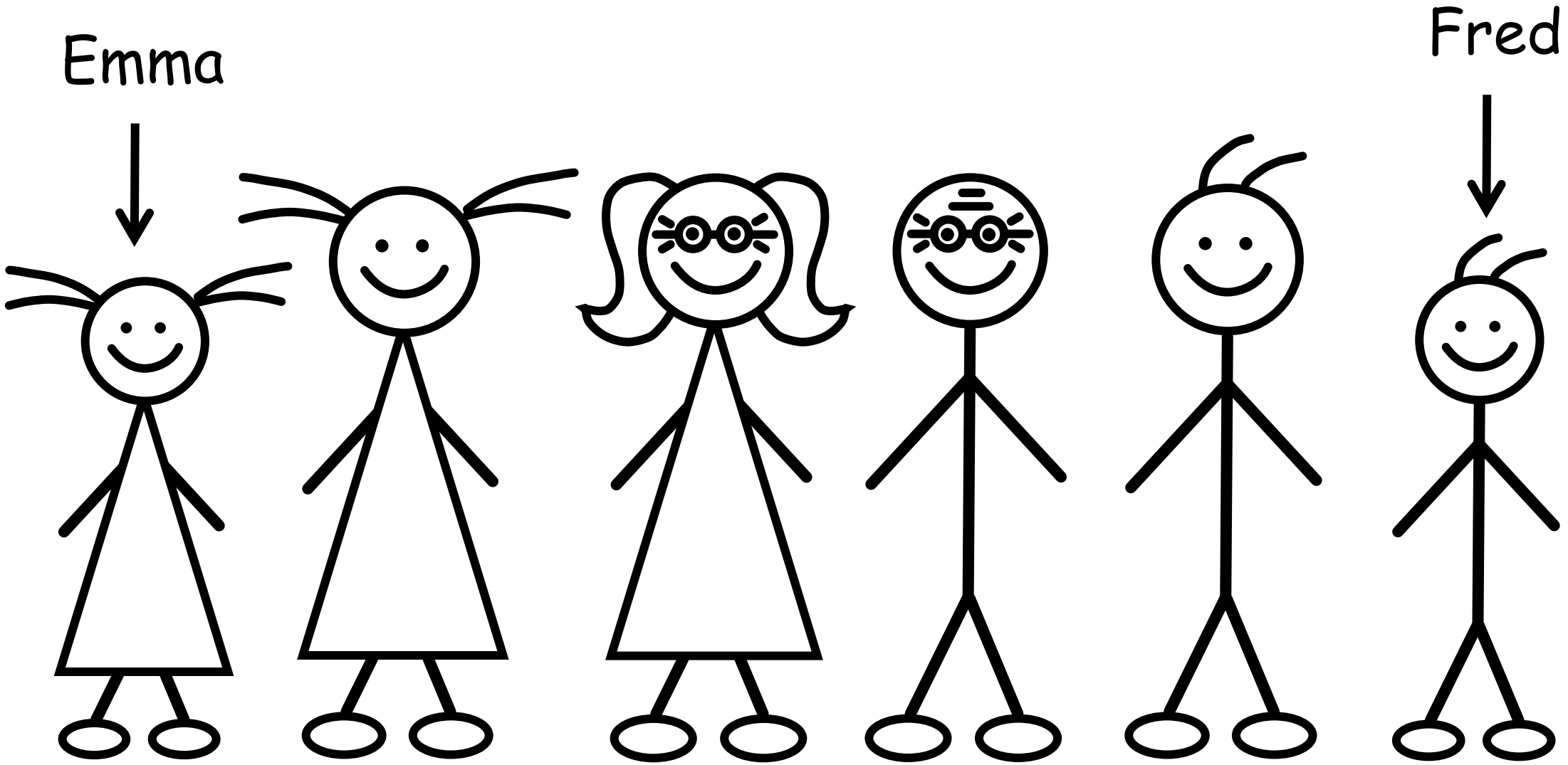
Chapter 3:
"AI Ethics of the Future"

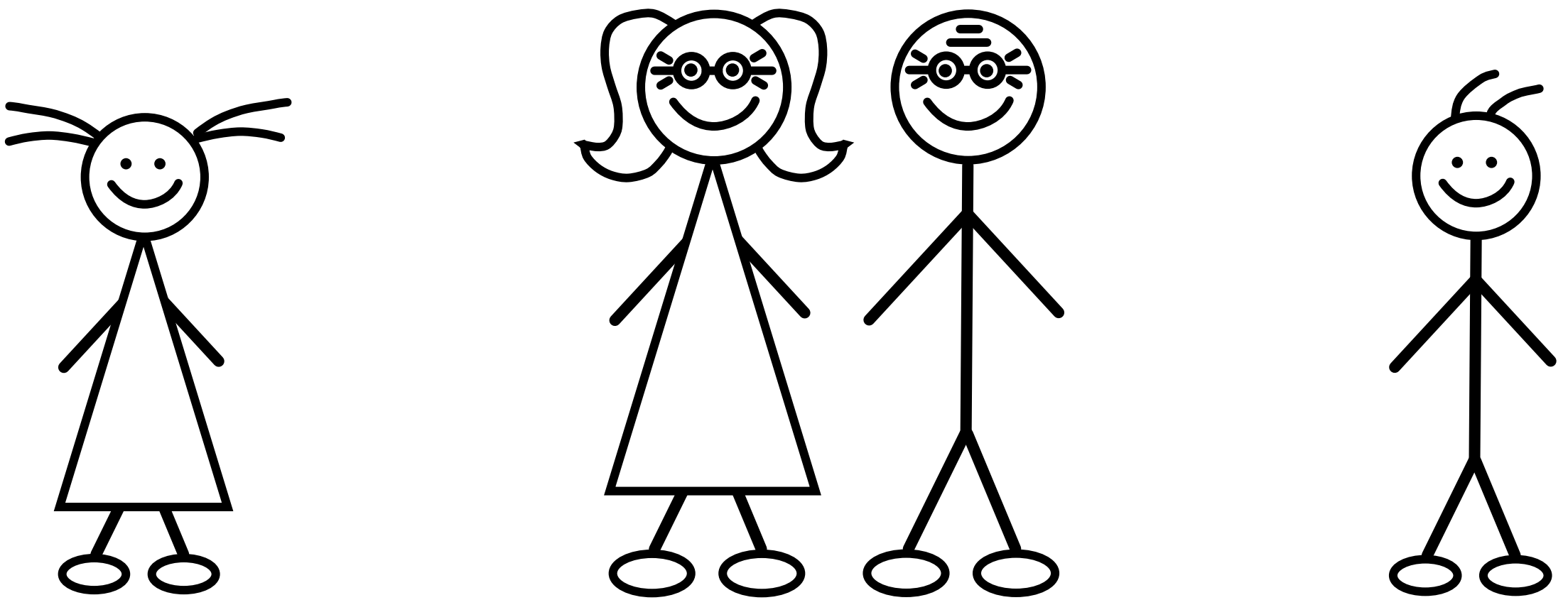


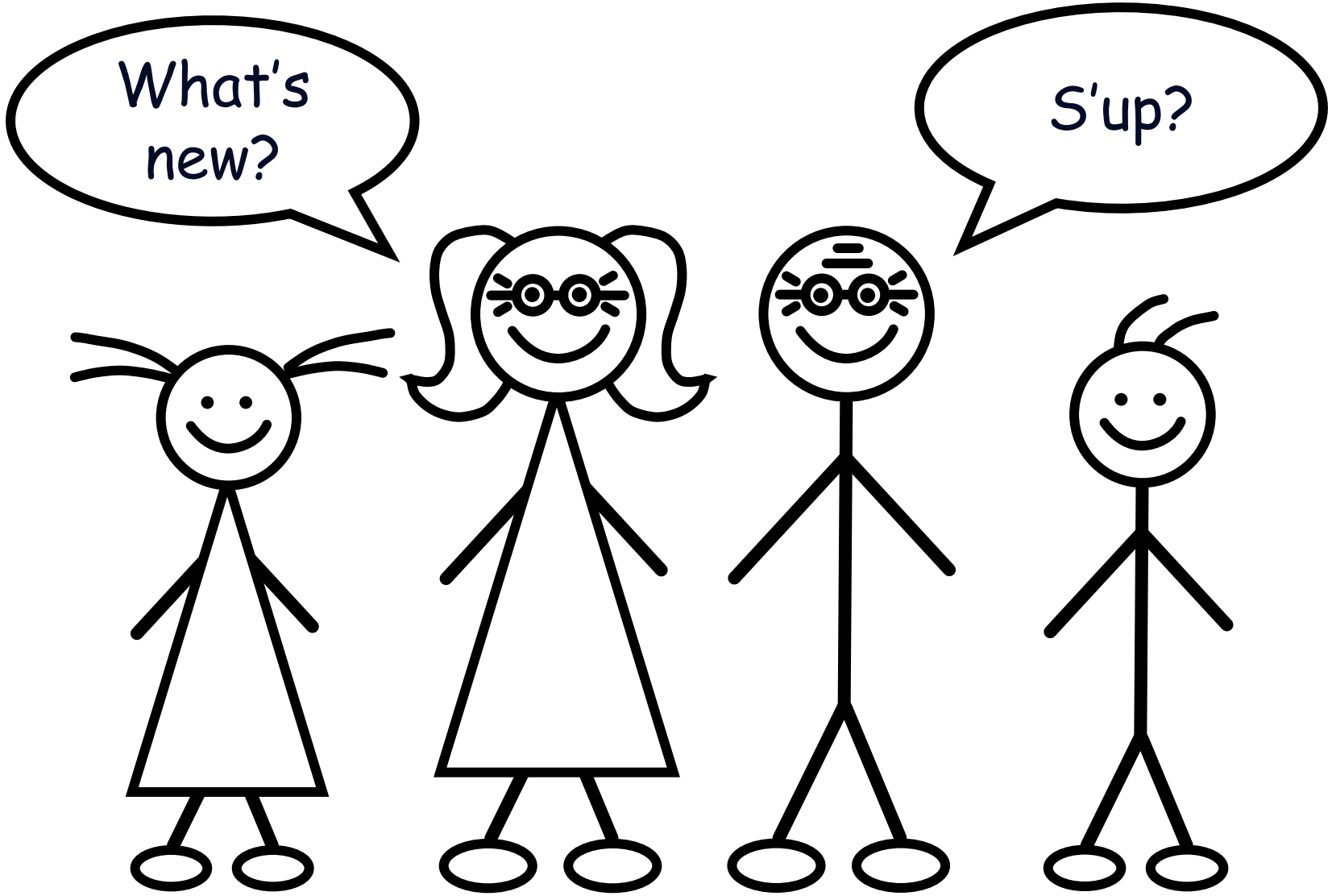


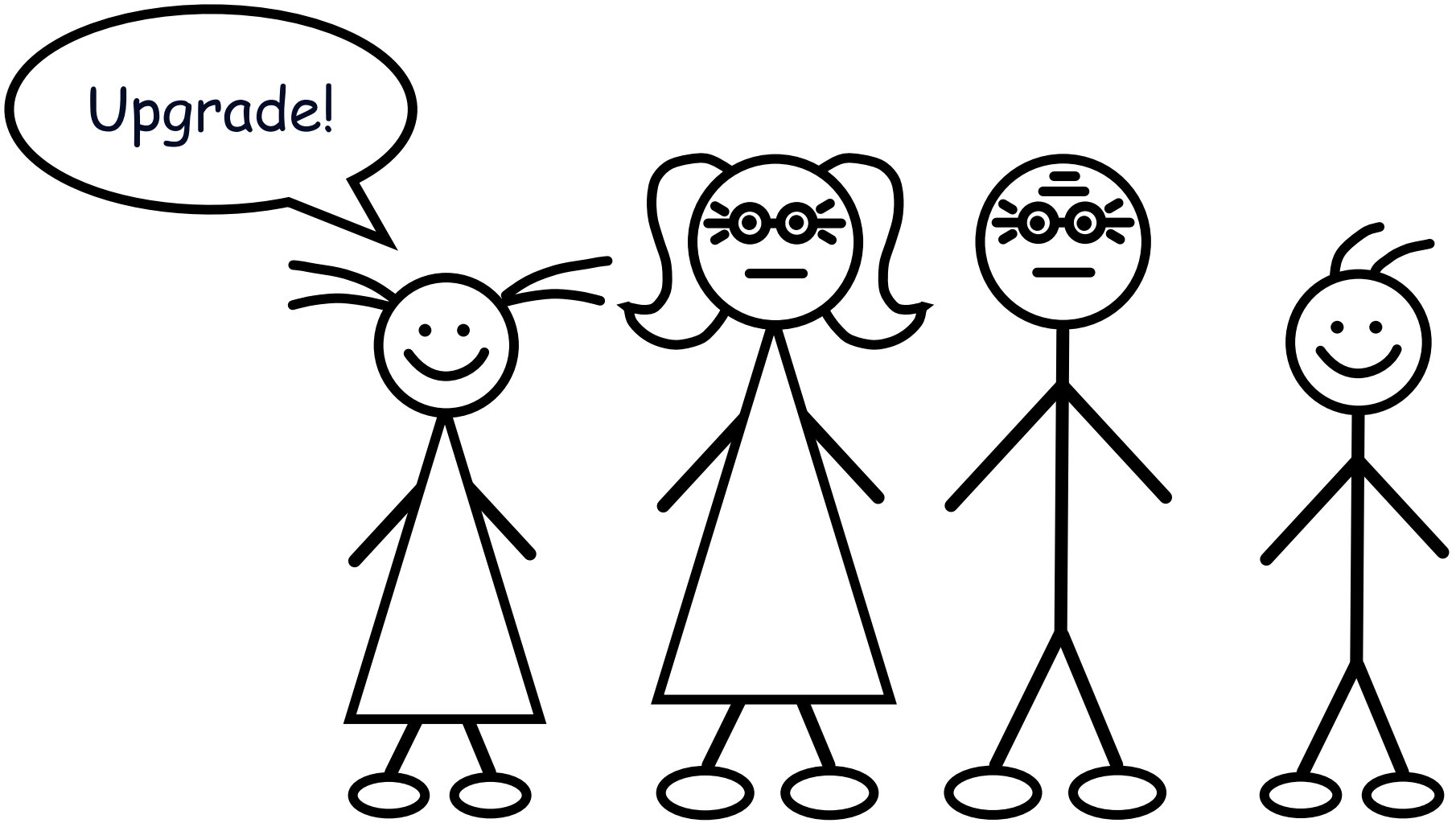


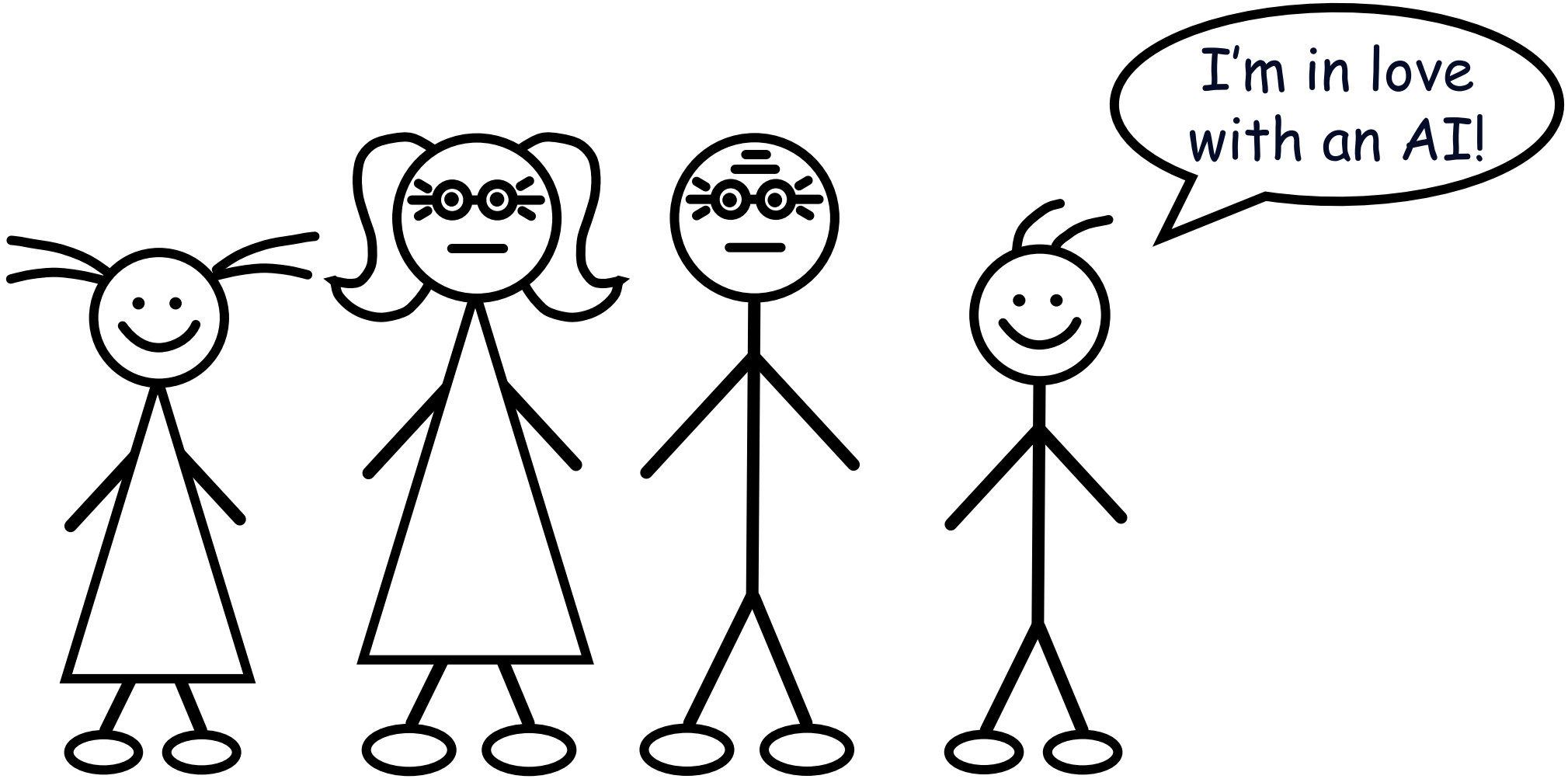

Emma

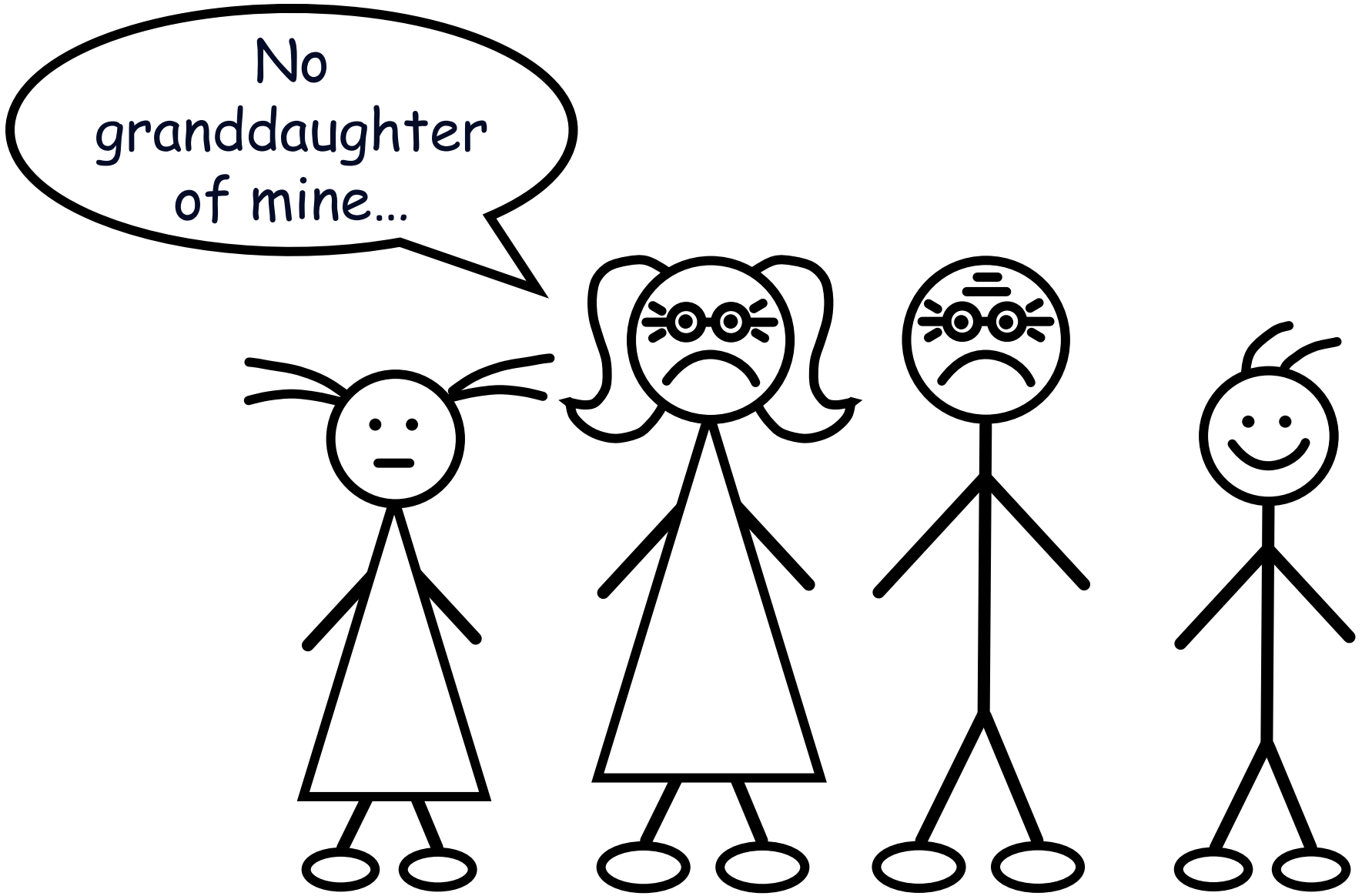

Emma
Fred

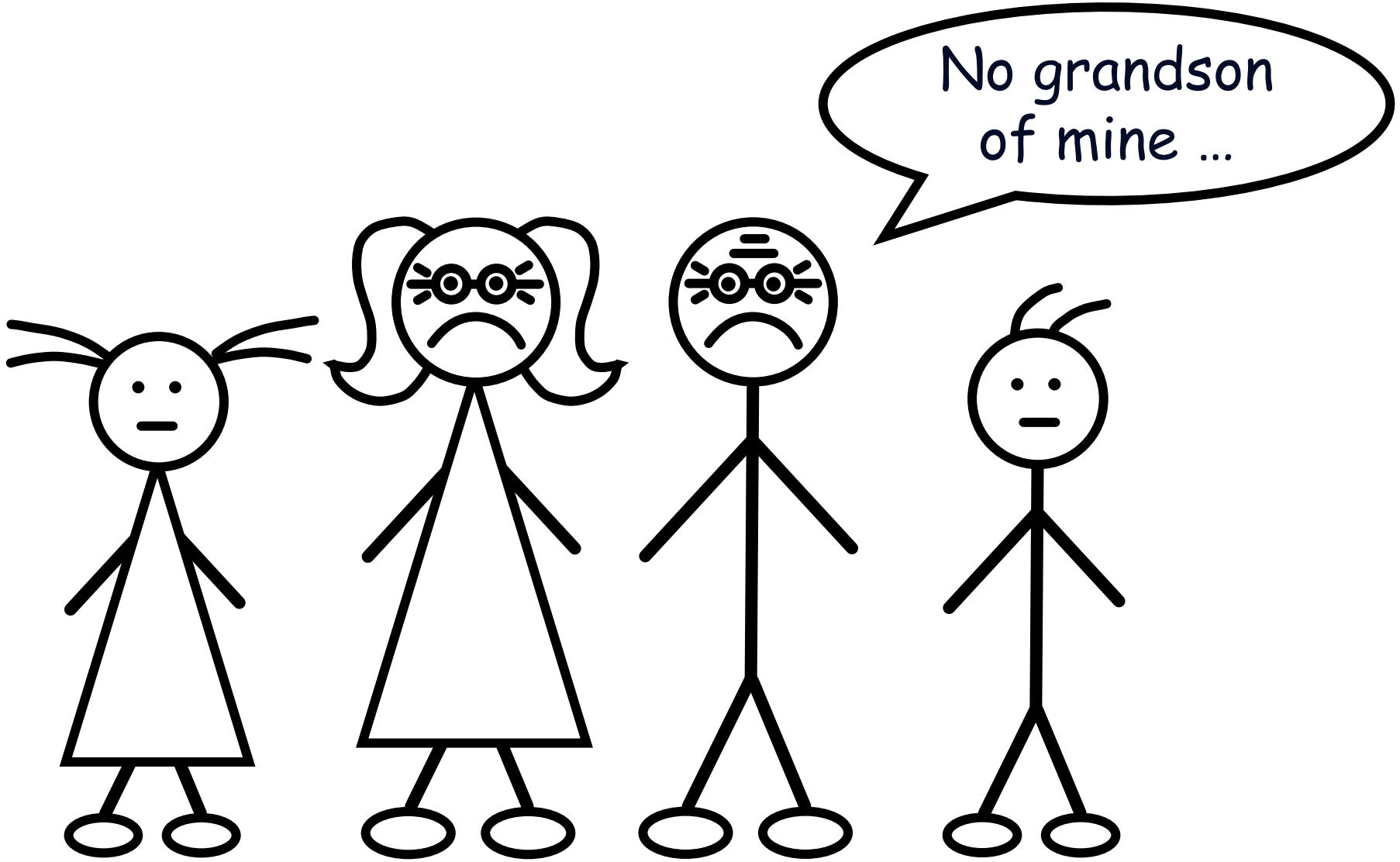


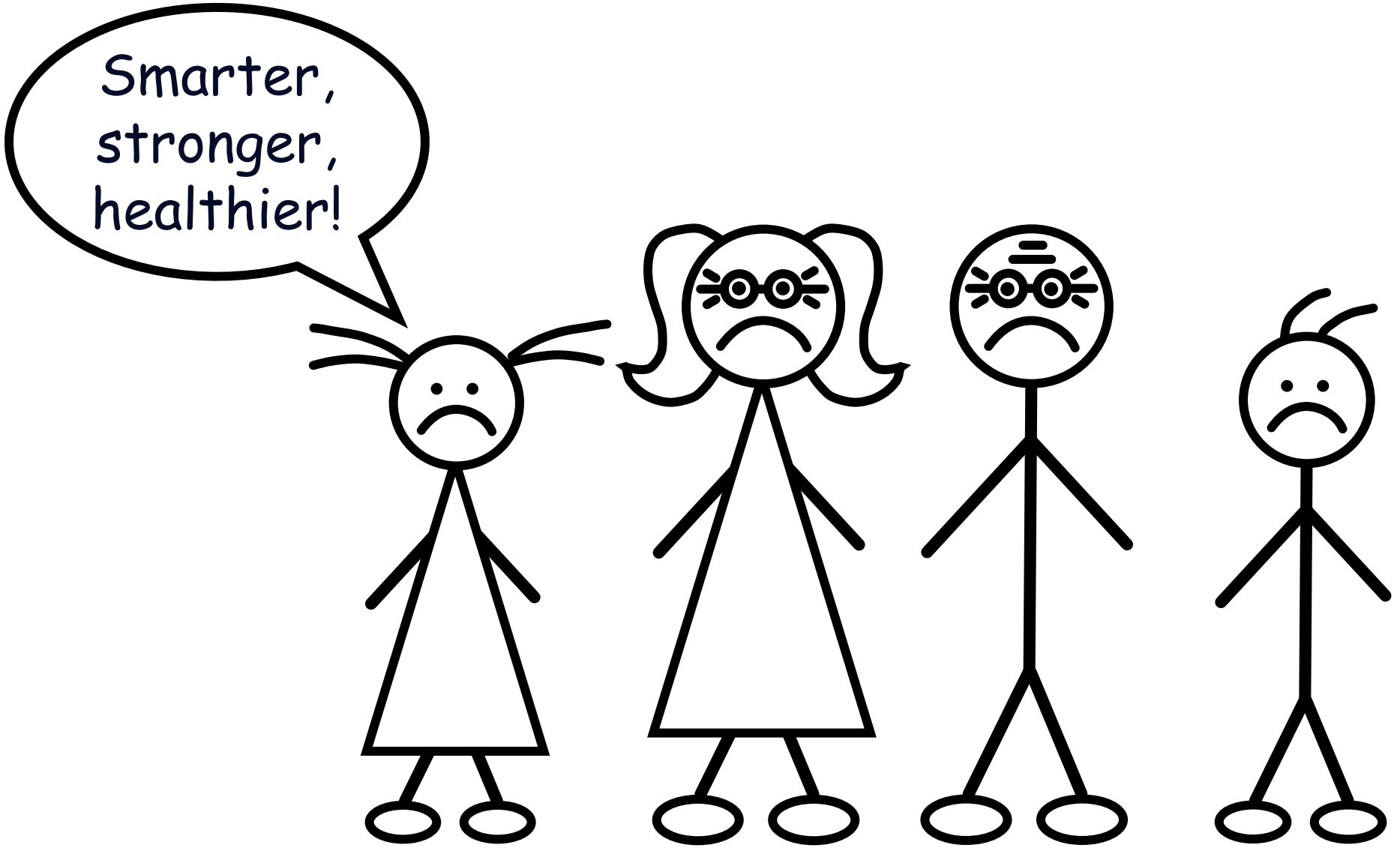

What's new?
S'up?

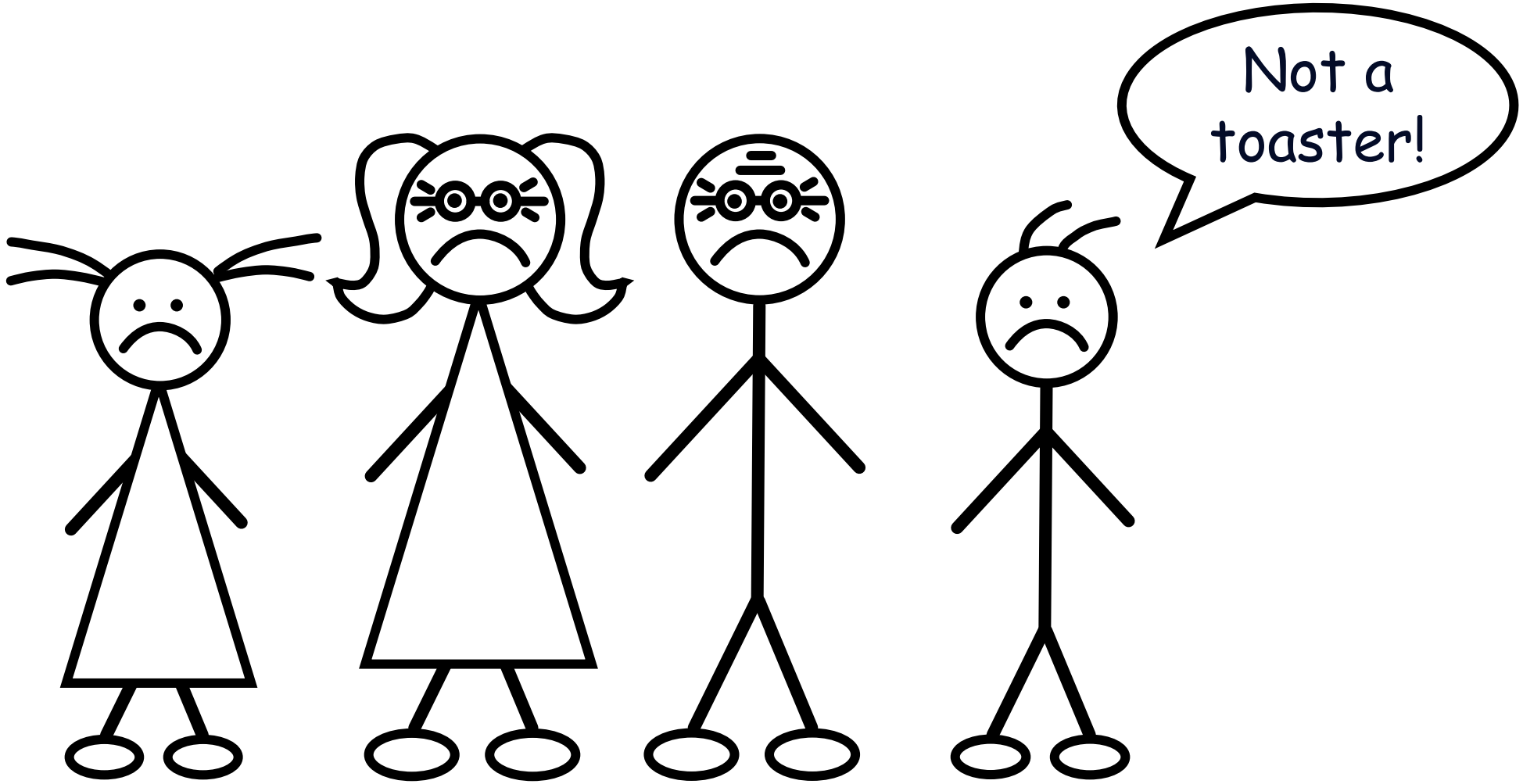

Upgrade!

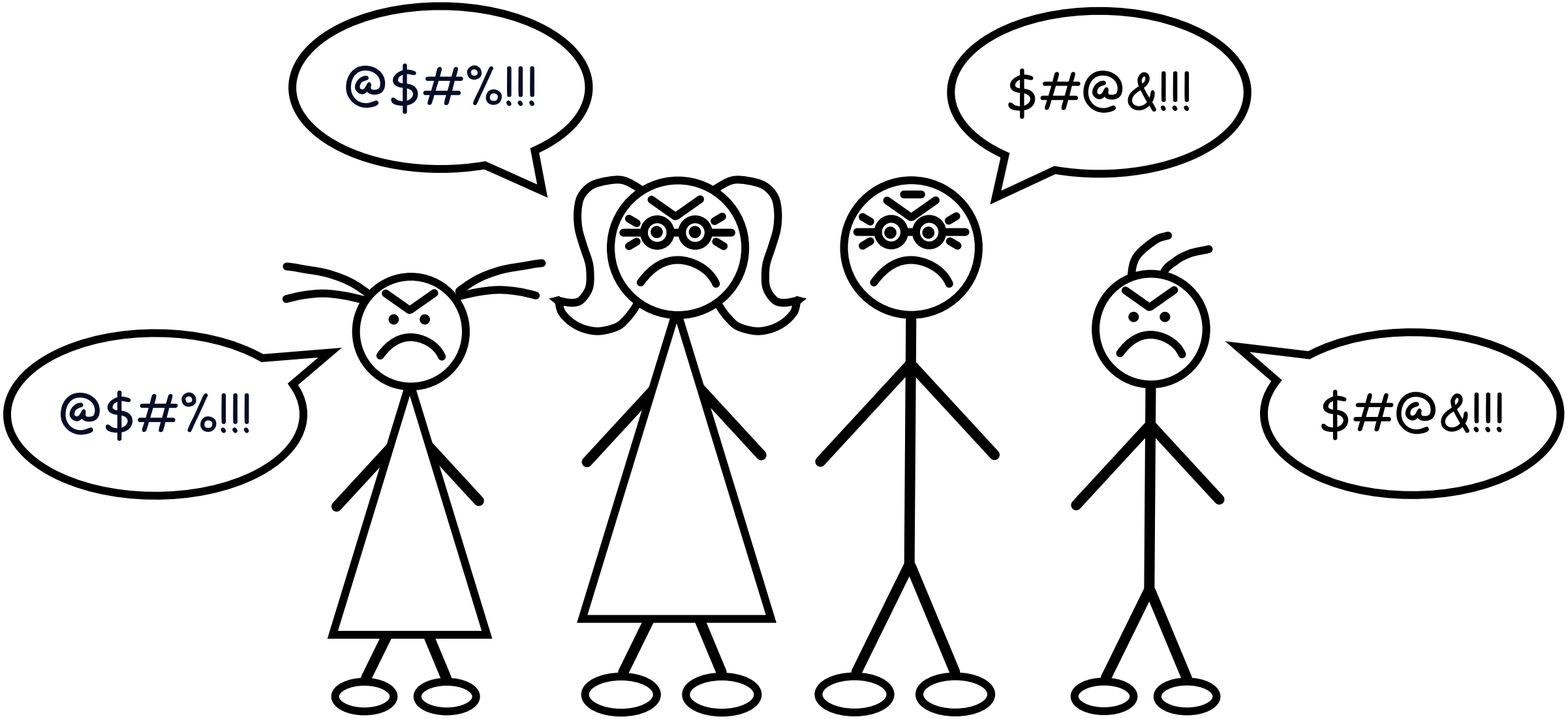

I'm in love with an AI!

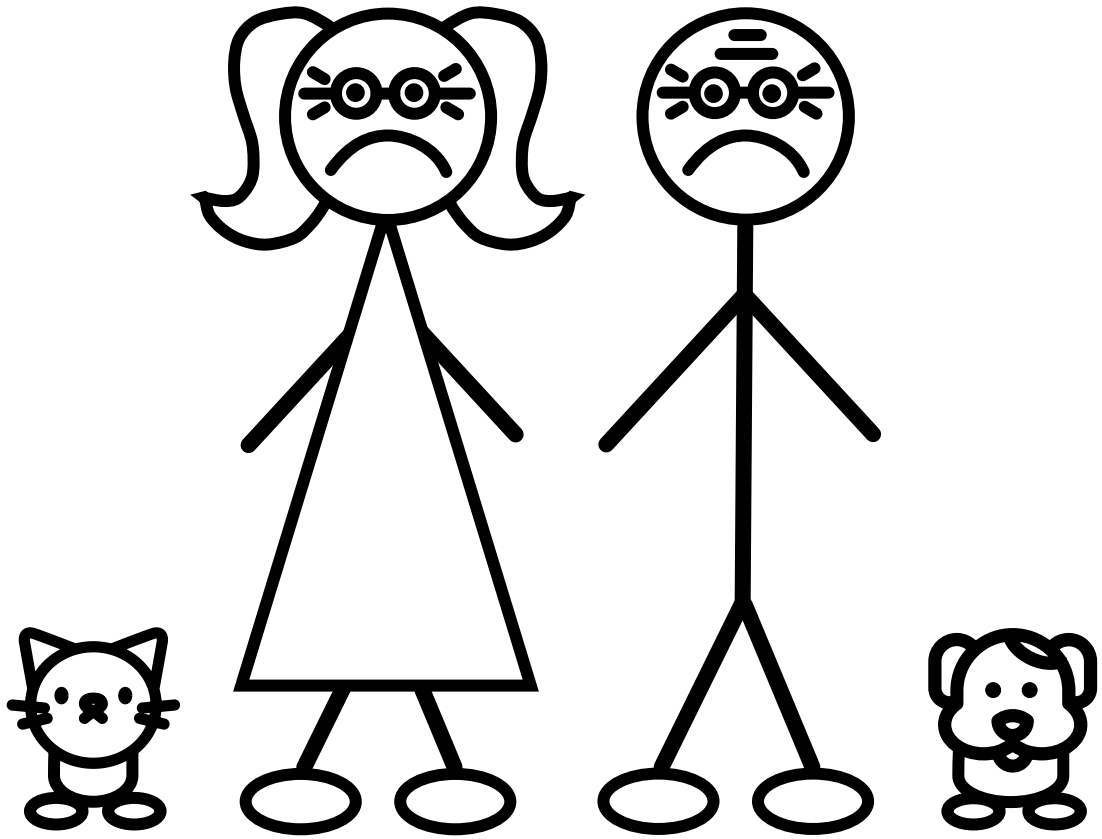

No granddaughter of mine…

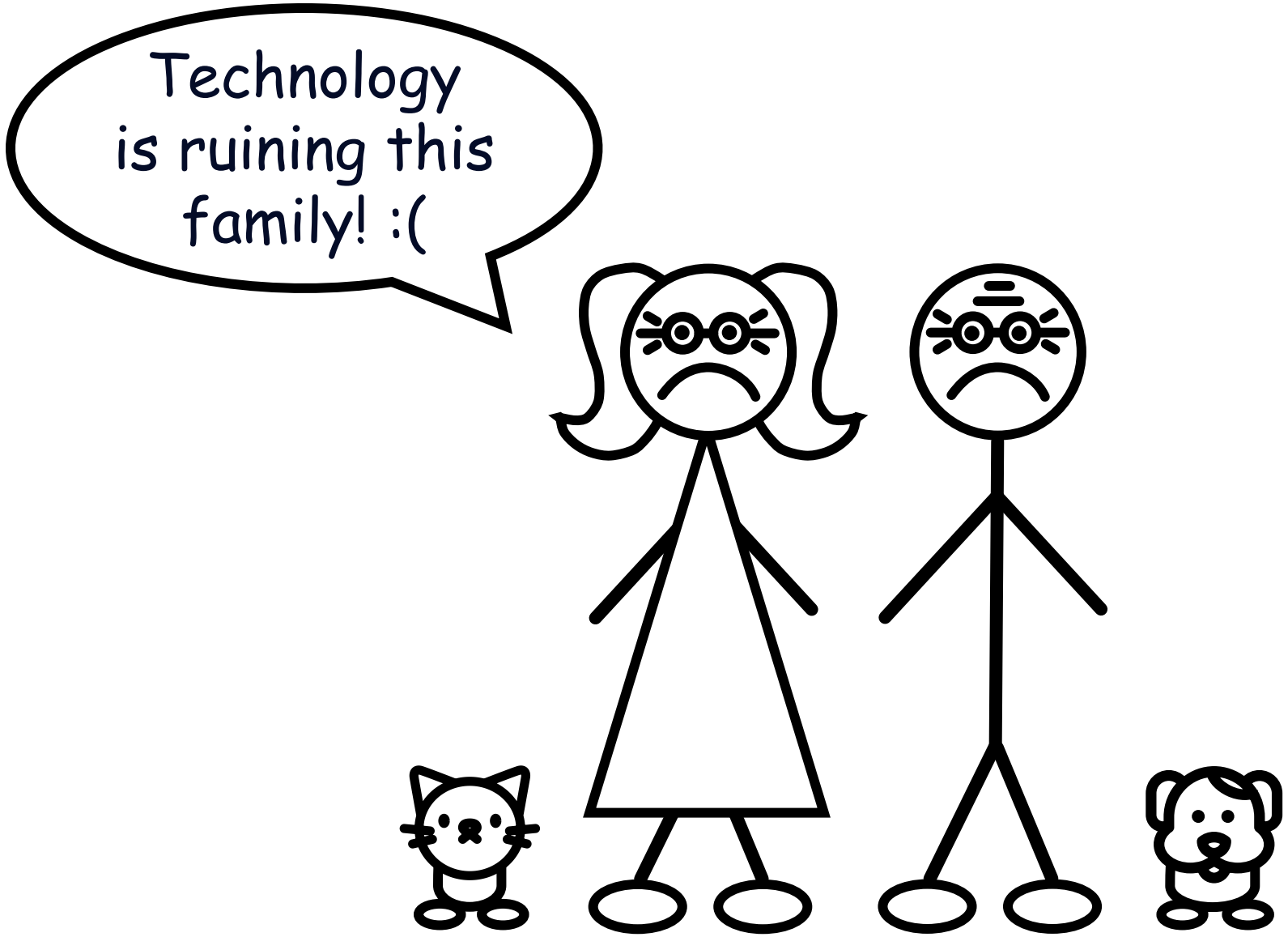

No grandson of mine ...

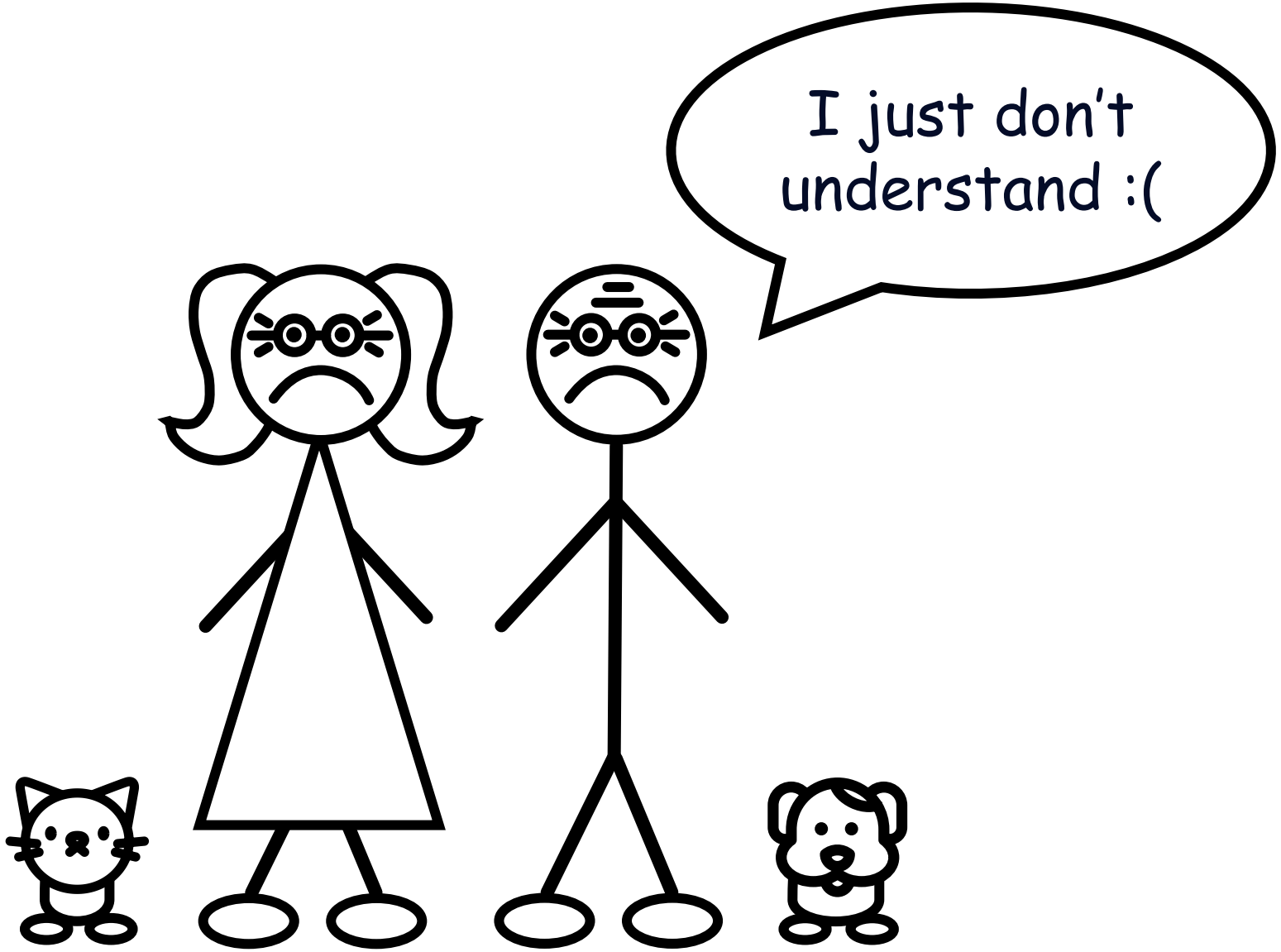

Smarter,
stronger,
healthier!

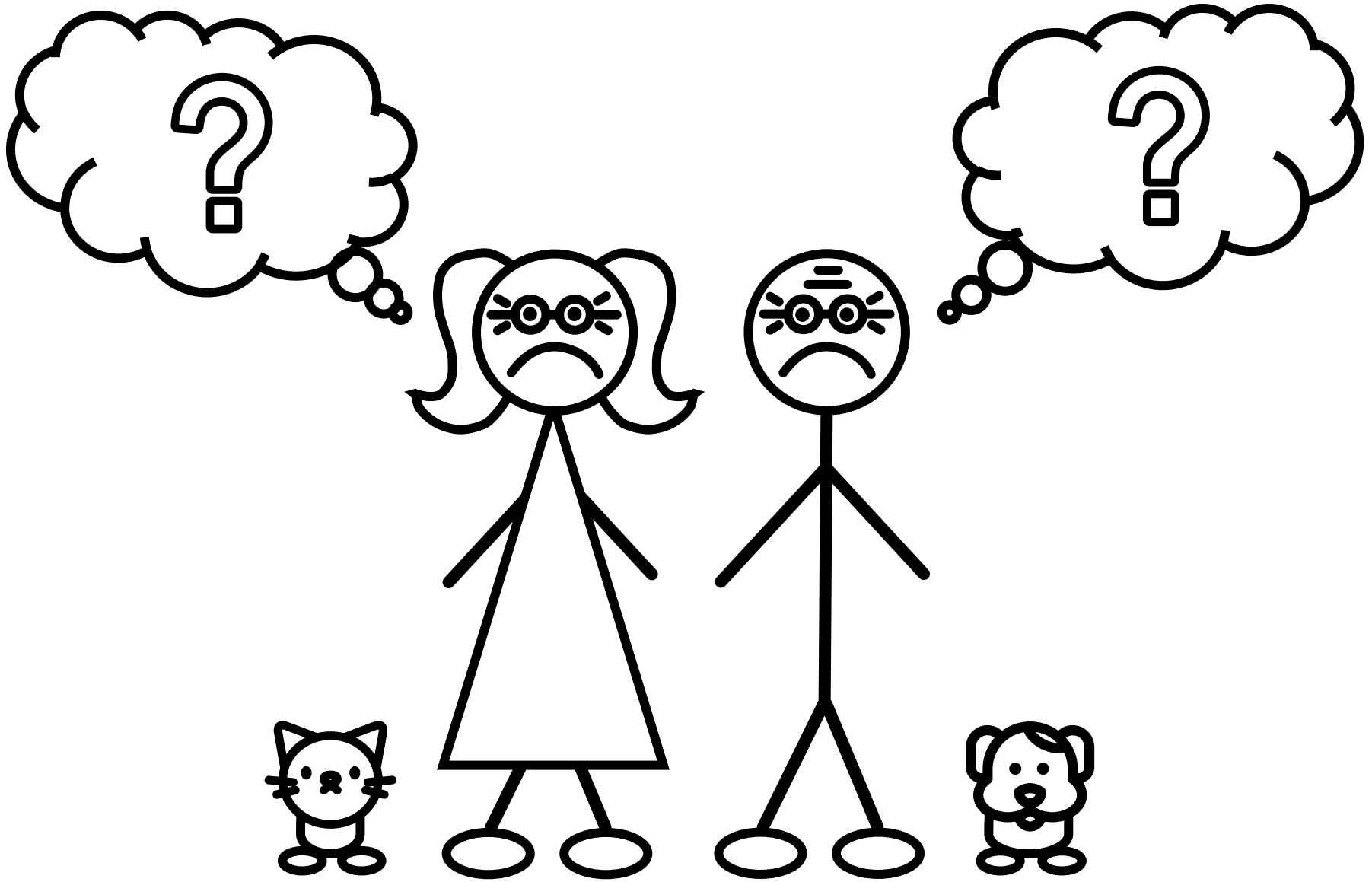

Not a toaster!






Technology is ruining this family! :(


I just don't understand :(



Benefits

Artificial General Intelligence (AGI)

Improve human health

Solve our biggest problems

Problems

Humans become obsolete

Inspired by CGP Grey

Horse vs.
the tractor

Humans become obsolete

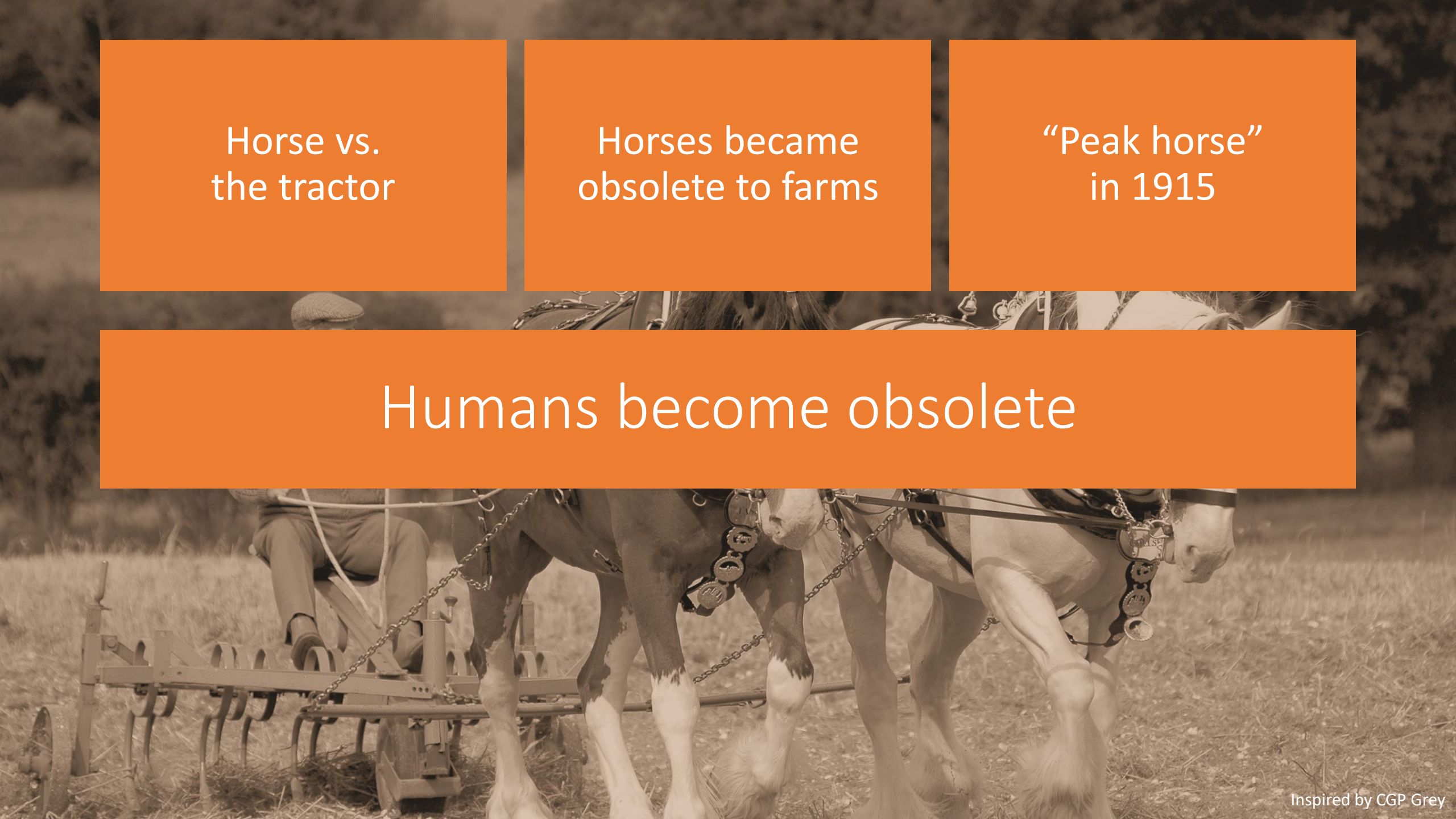
Inspired by CGP Grey

Horse vs.
the tractor

Horses became
obsolete to farms

Humans become obsolete

Inspired by CGP Grey



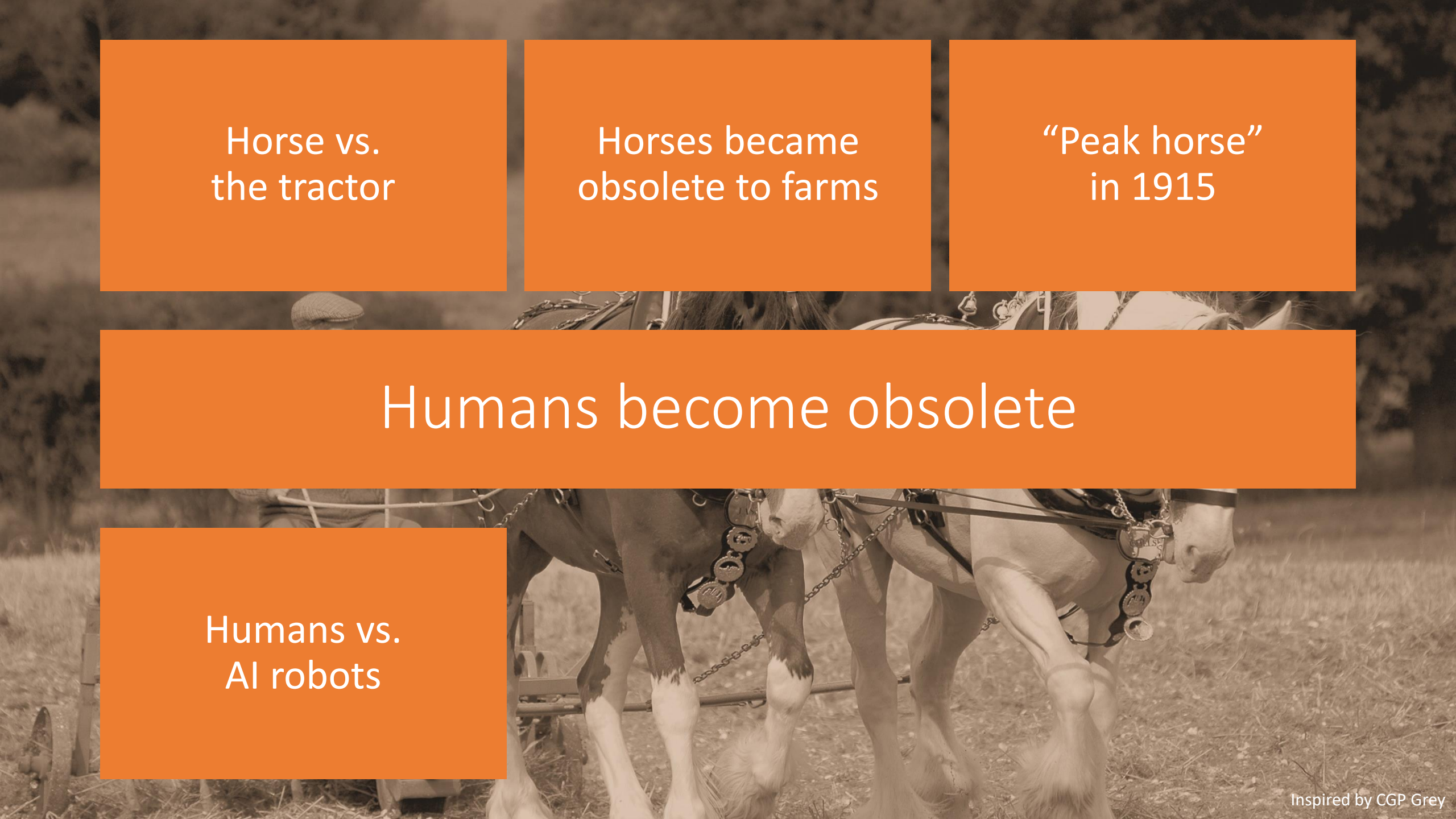
Horse vs.
the tractor

Horses became
obsolete to farms

“Peak horse”
in 1915

Humans become obsolete

Inspired by CGP Grey



Horse vs.
the tractor

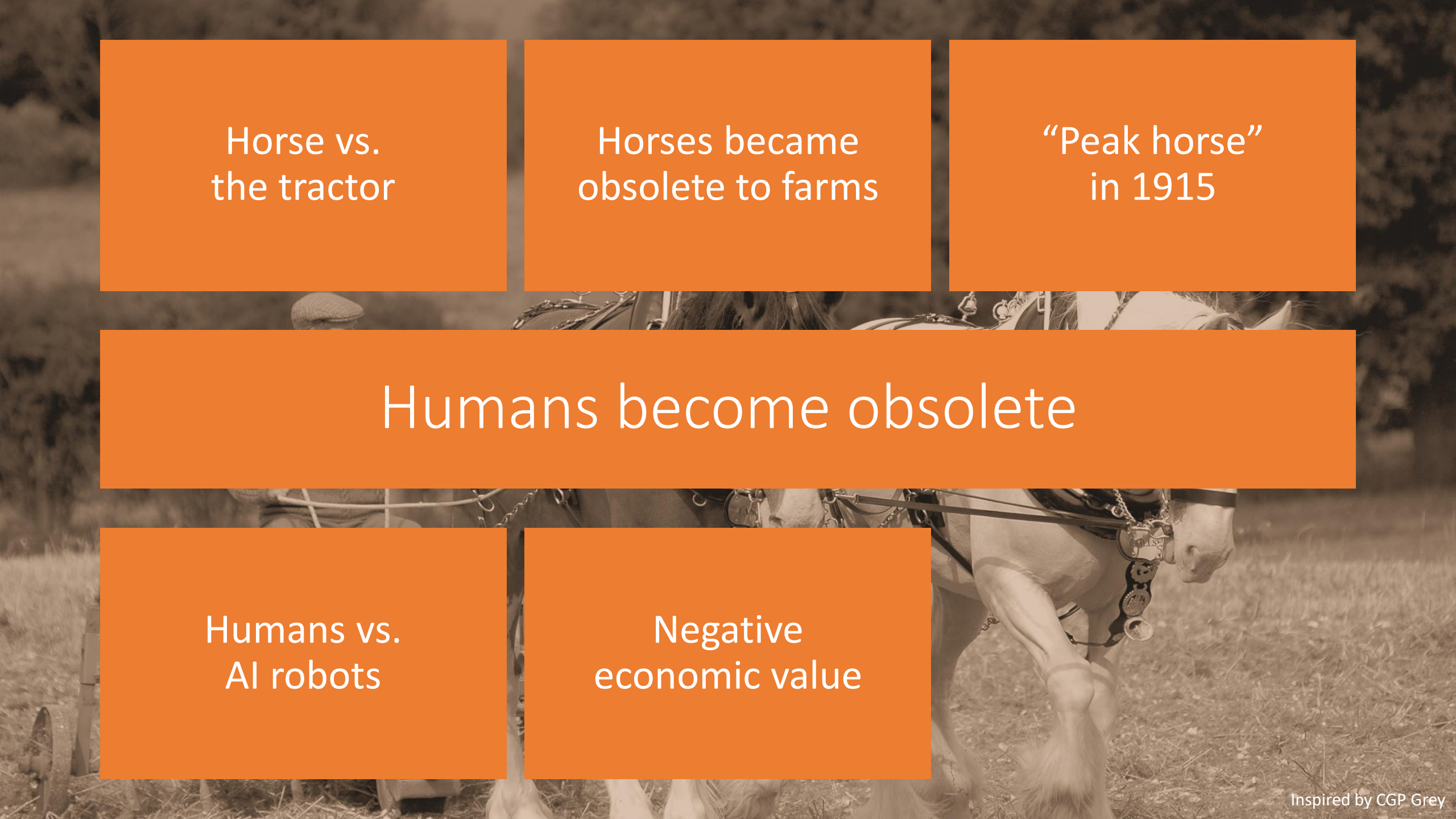
Horses became
obsolete to farms

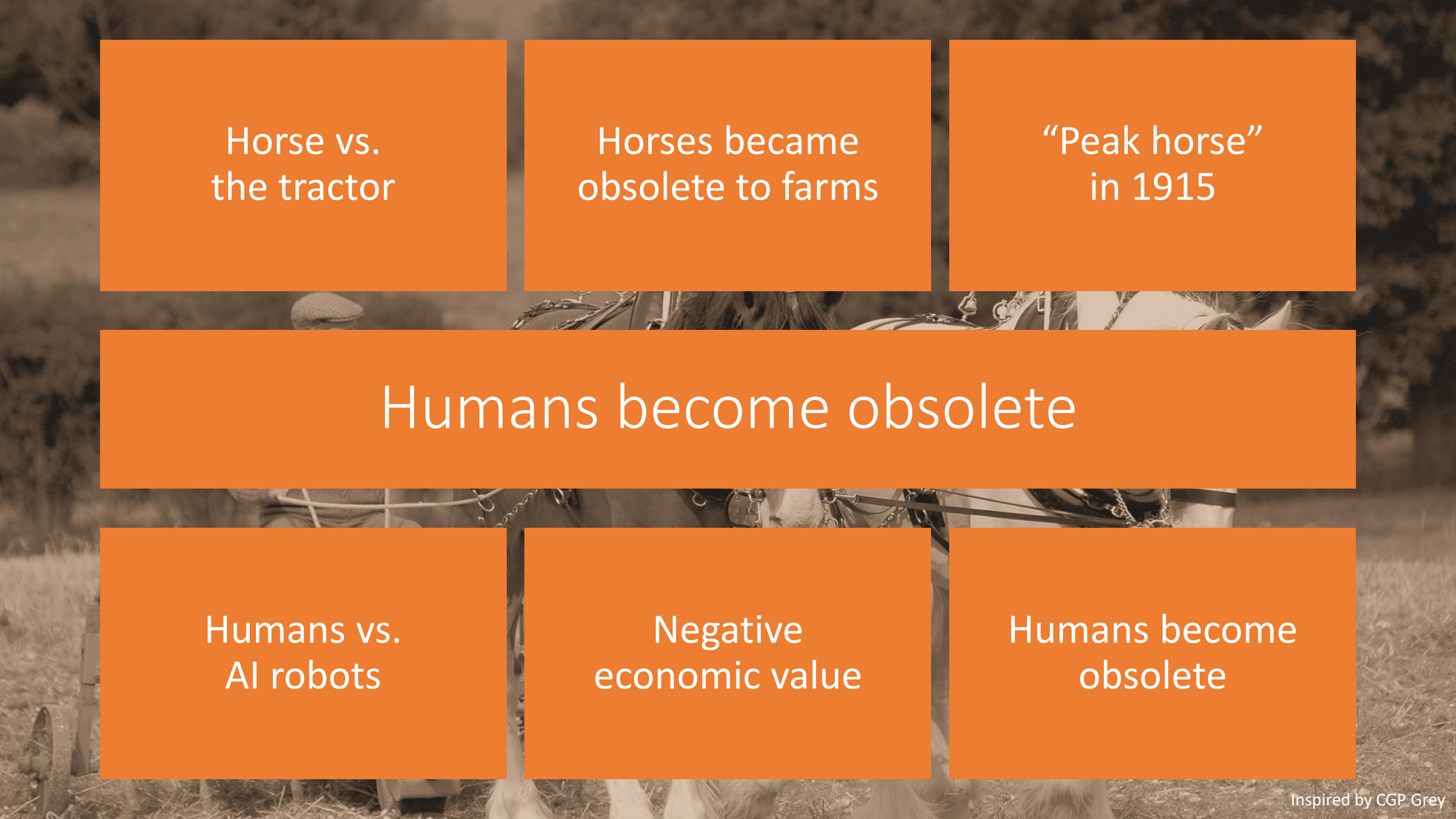
“Peak horse”
in 1915

Humans become obsolete

Humans vs.
AI robots

Inspired by CGP Grey


Horse vs.
the tractor
Horses became
obsolete to farms
“Peak horse”
in 1915
Humans become obsolete
Humans vs.
AI robots
Negative
economic value
Inspired by CGP Grey



Horse vs.
the tractor

Horses became
obsolete to farms

“Peak horse”
in 1915

Humans become obsolete

Humans vs.
AI robots

Negative
economic value

Humans become
obsolete

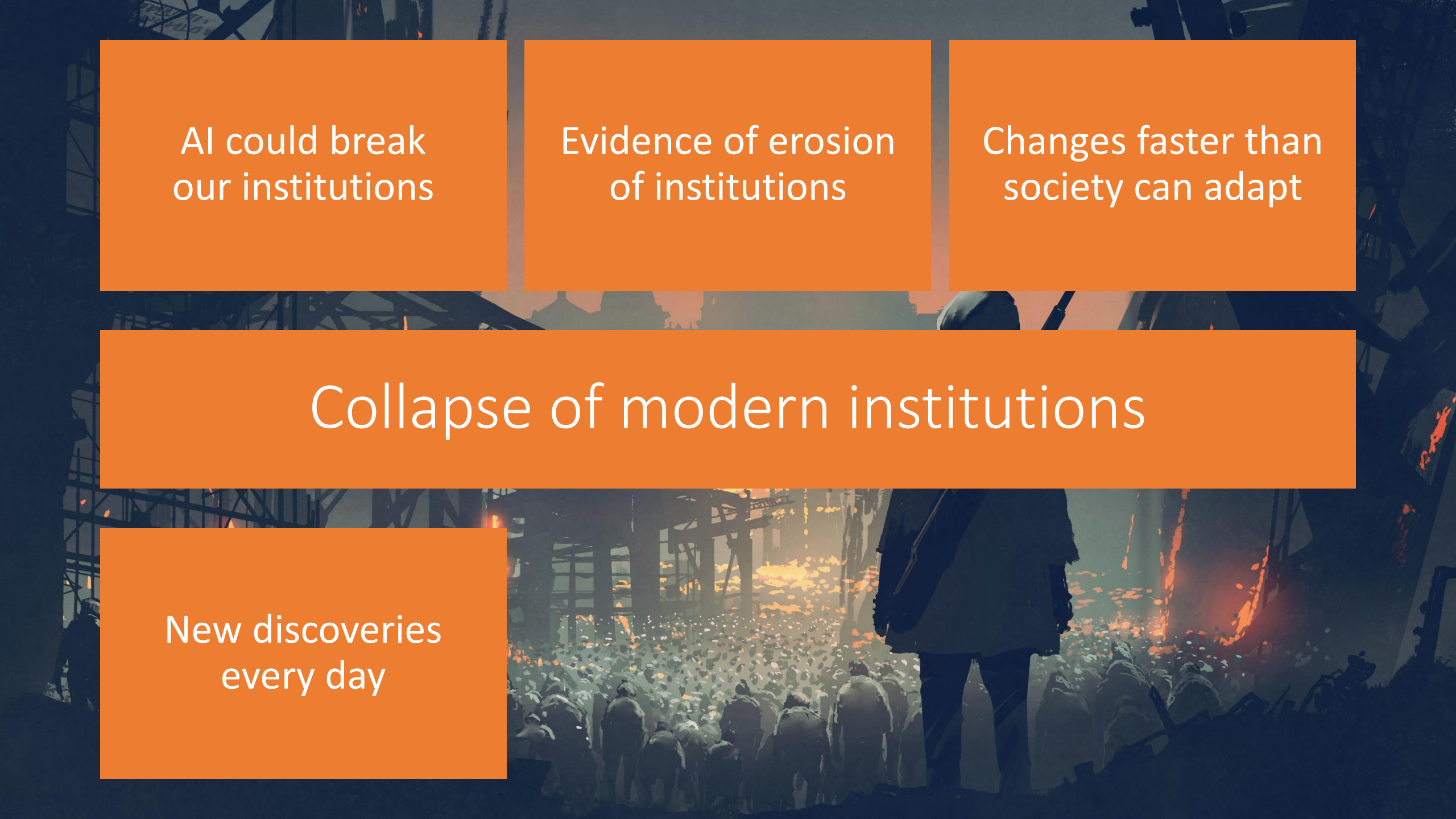
Inspired by CGP Grey

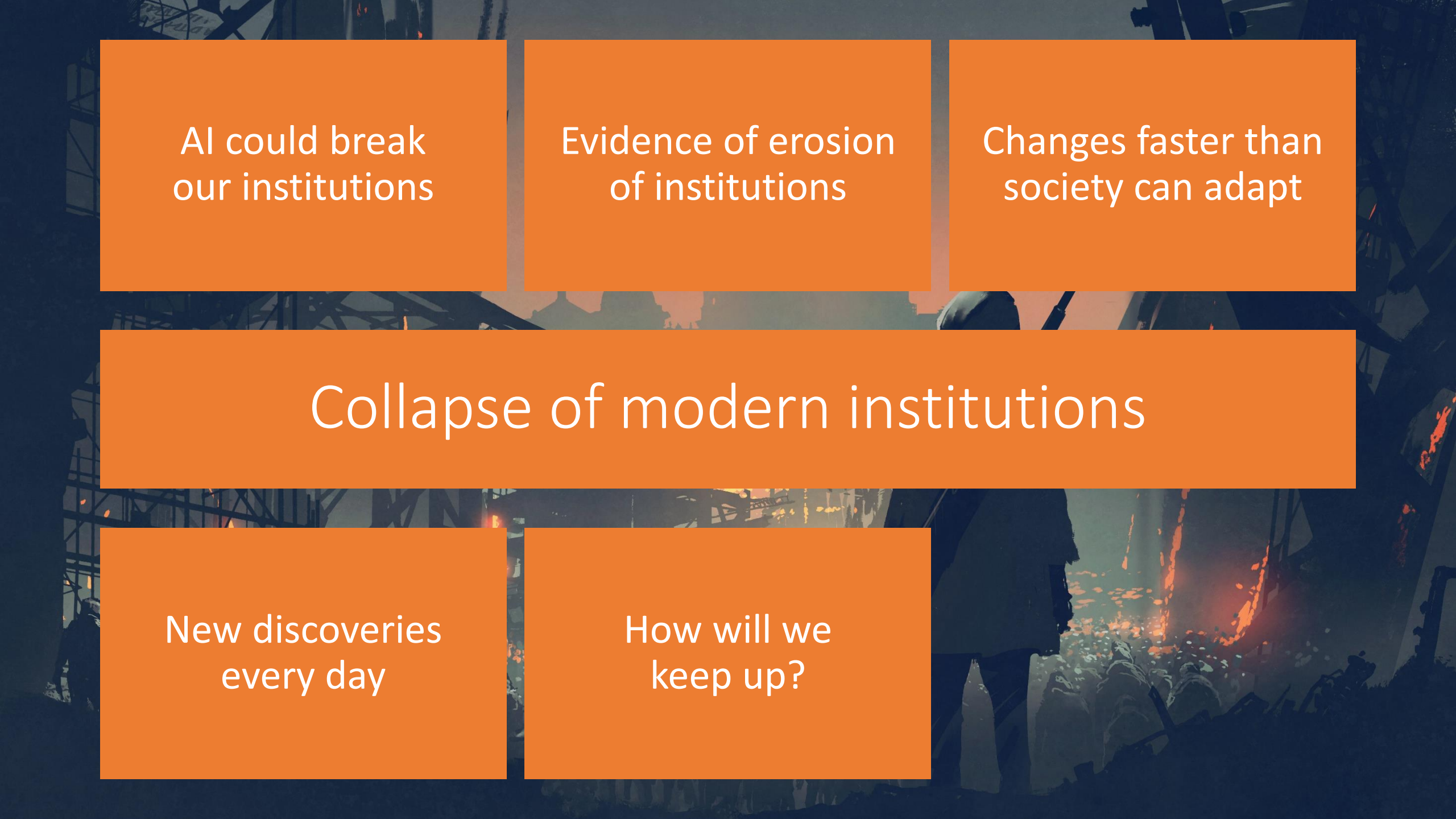
Collapse of modern institutions

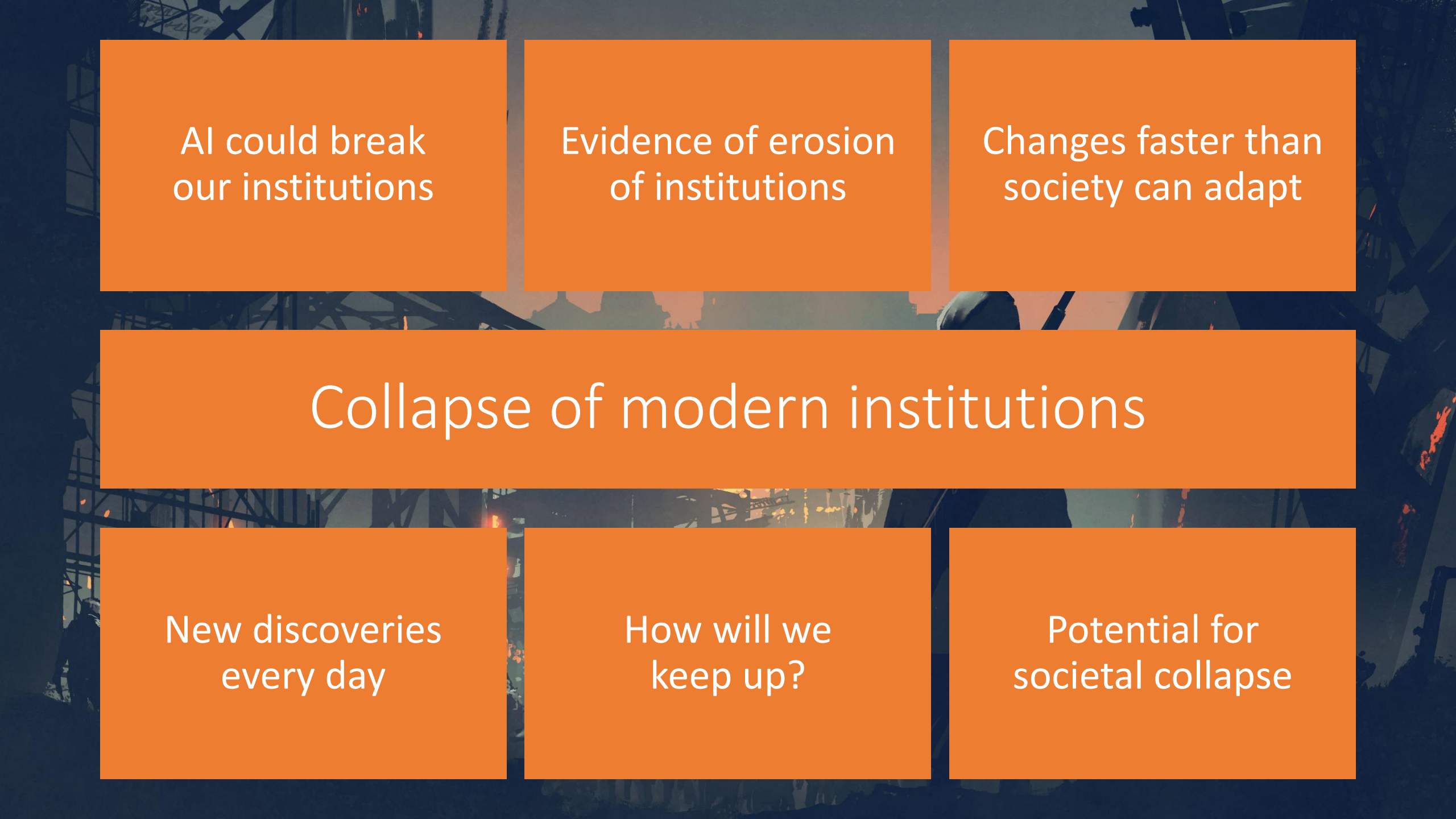
AI could break our institutions
Collapse of modern institutions

AI could break our institutions
Evidence of erosion of institutions
Collapse of modern institutions

AI could break our institutions
Evidence of erosion of institutions
Changes faster than society can adapt
Collapse of modern institutions


AI could break our institutions
Evidence of erosion of institutions
Changes faster than society can adapt
Collapse of modern institutions
New discoveries every day


AI could break our institutions
Evidence of erosion of institutions
Changes faster than society can adapt
Collapse of modern institutions
New discoveries every day
How will we keep up?


AI could break our institutions
Evidence of erosion of institutions
Changes faster than society can adapt
Collapse of modern institutions
New discoveries every day
How will we keep up?
Potential for societal collapse

Dystopian future

The return of authoritarianism

Dystopian future

The return of authoritarianism

The return of communism

Dystopian future

The return of authoritarianism

The return of communism

The return of fascism

Dystopian future

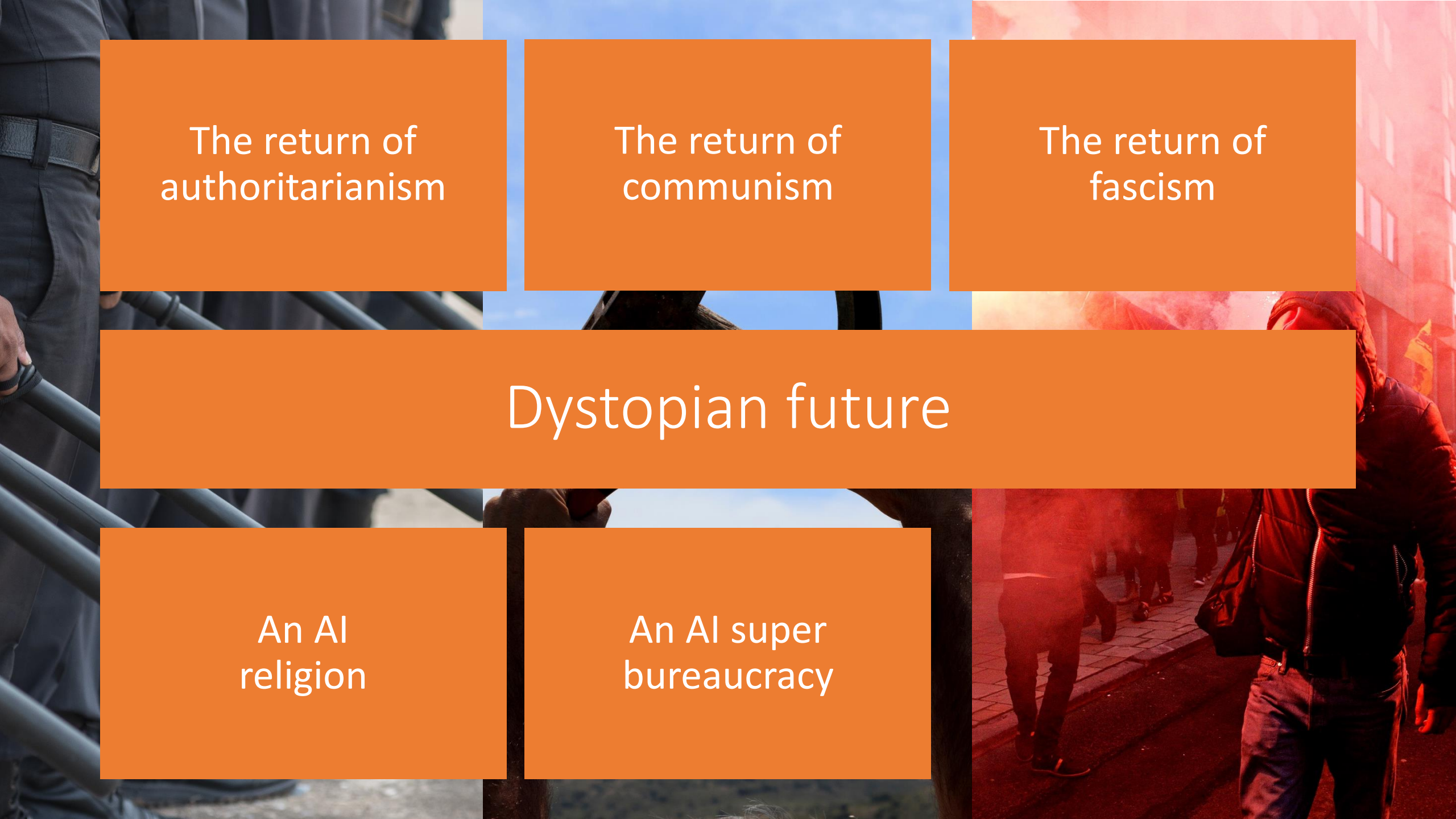
The return of authoritarianism

The return of communism

The return of fascism

Dystopian future

An AI religion



The return of authoritarianism

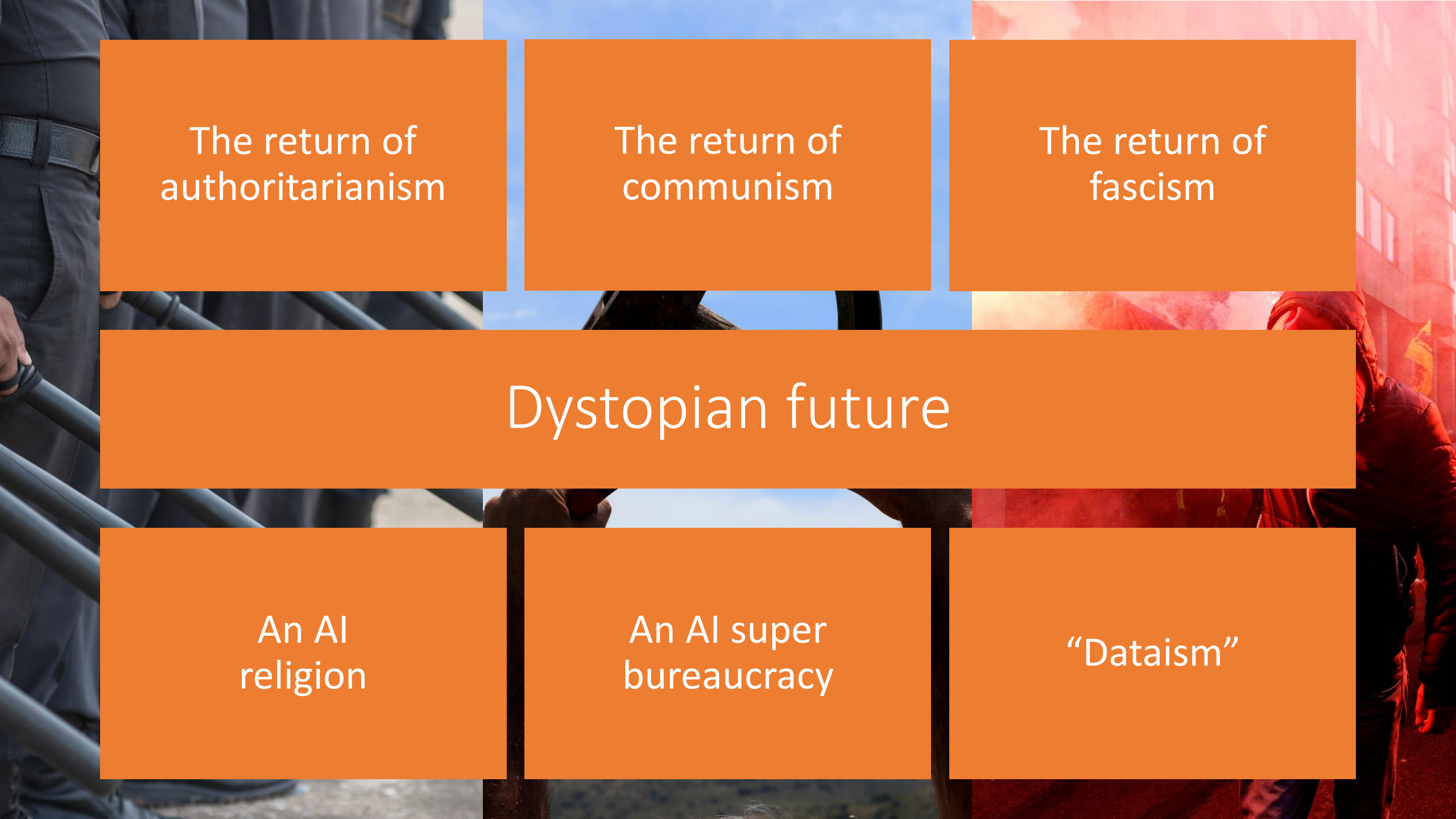
The return of communism

The return of fascism

Dystopian future

An AI religion

An AI super bureaucracy



The return of authoritarianism

The return of communism

The return of fascism

Dystopian future

An AI religion

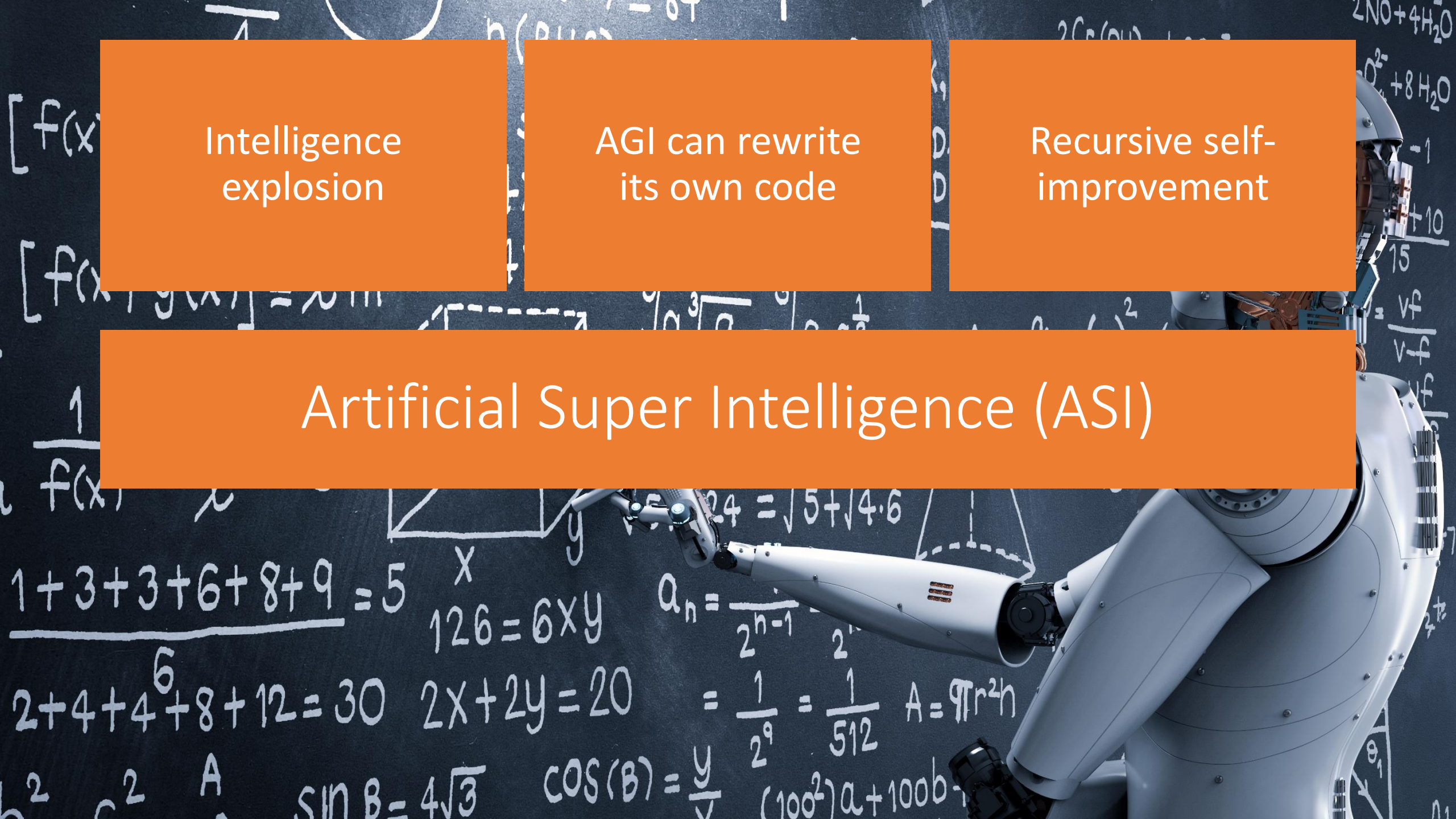
An AI super bureaucracy

“Dataism”

Artificial Super Intelligence (ASI)

Intelligence explosion
Artificial Super Intelligence (ASI)

Intelligence explosion
AGI can rewrite its own code
Artificial Super Intelligence (ASI)


Intelligence explosion
AGI can rewrite its own code
Recursive self-improvement
Artificial Super Intelligence (ASI)

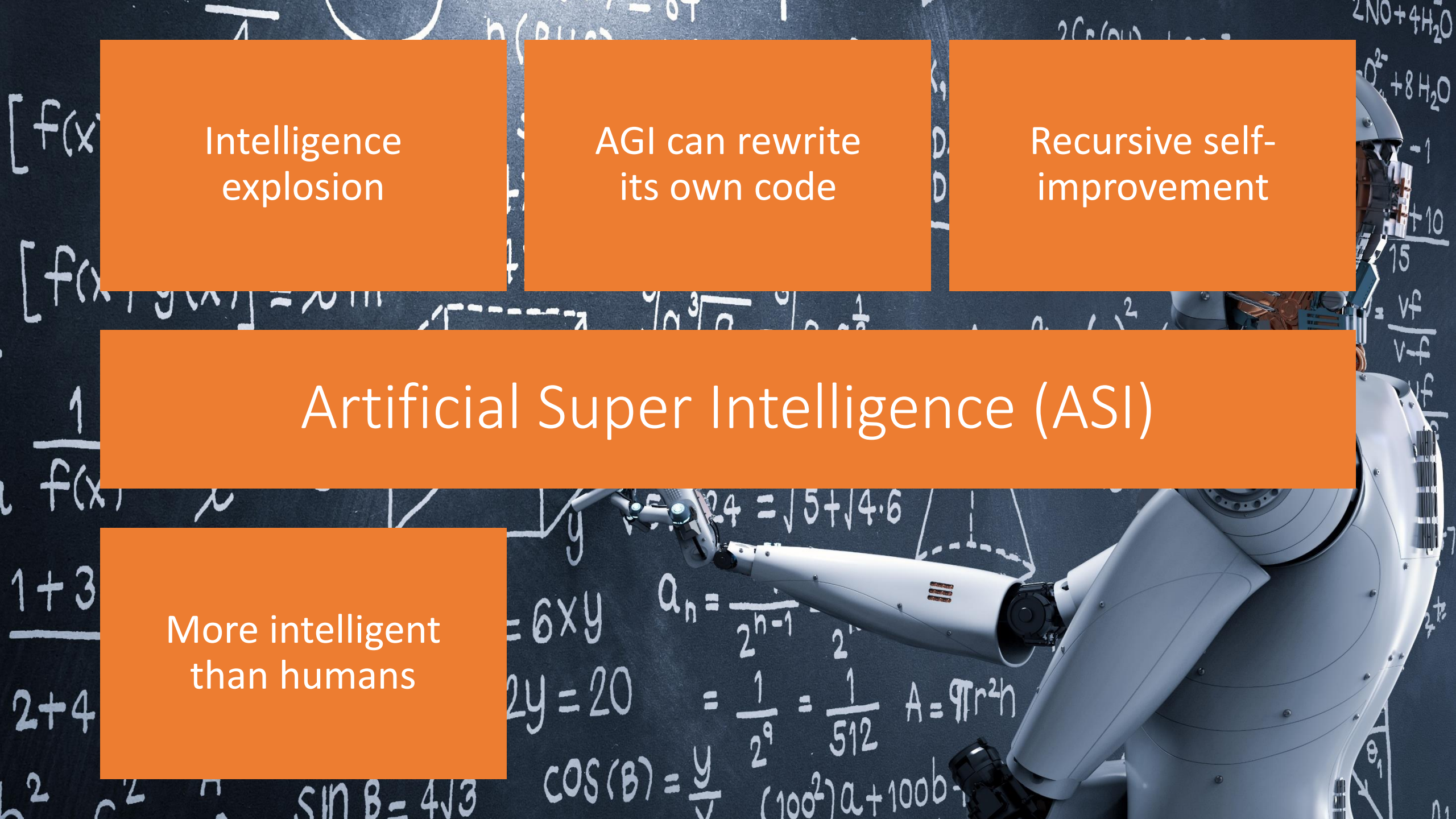
Intelligence explosion

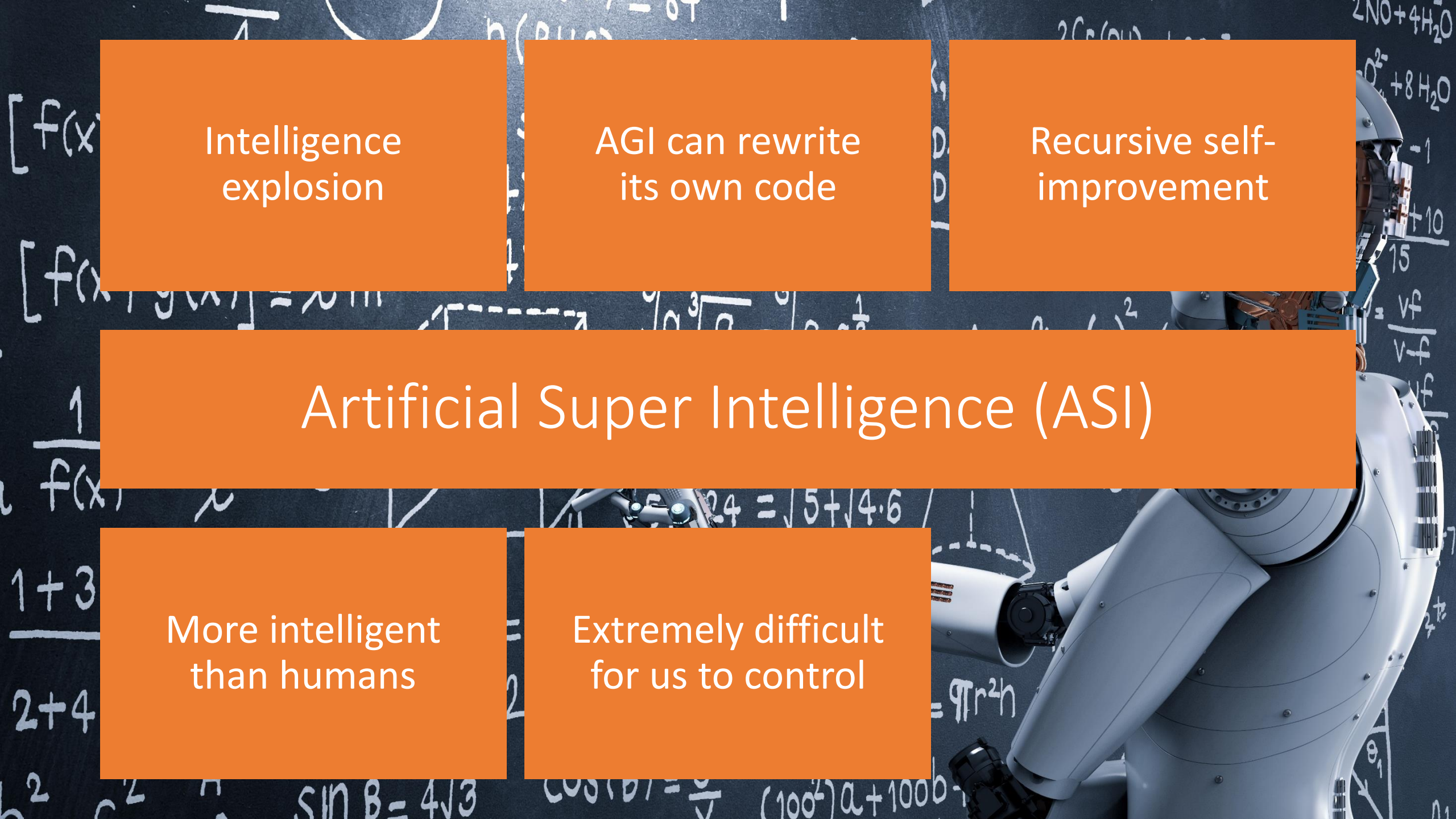
AGI can rewrite its own code

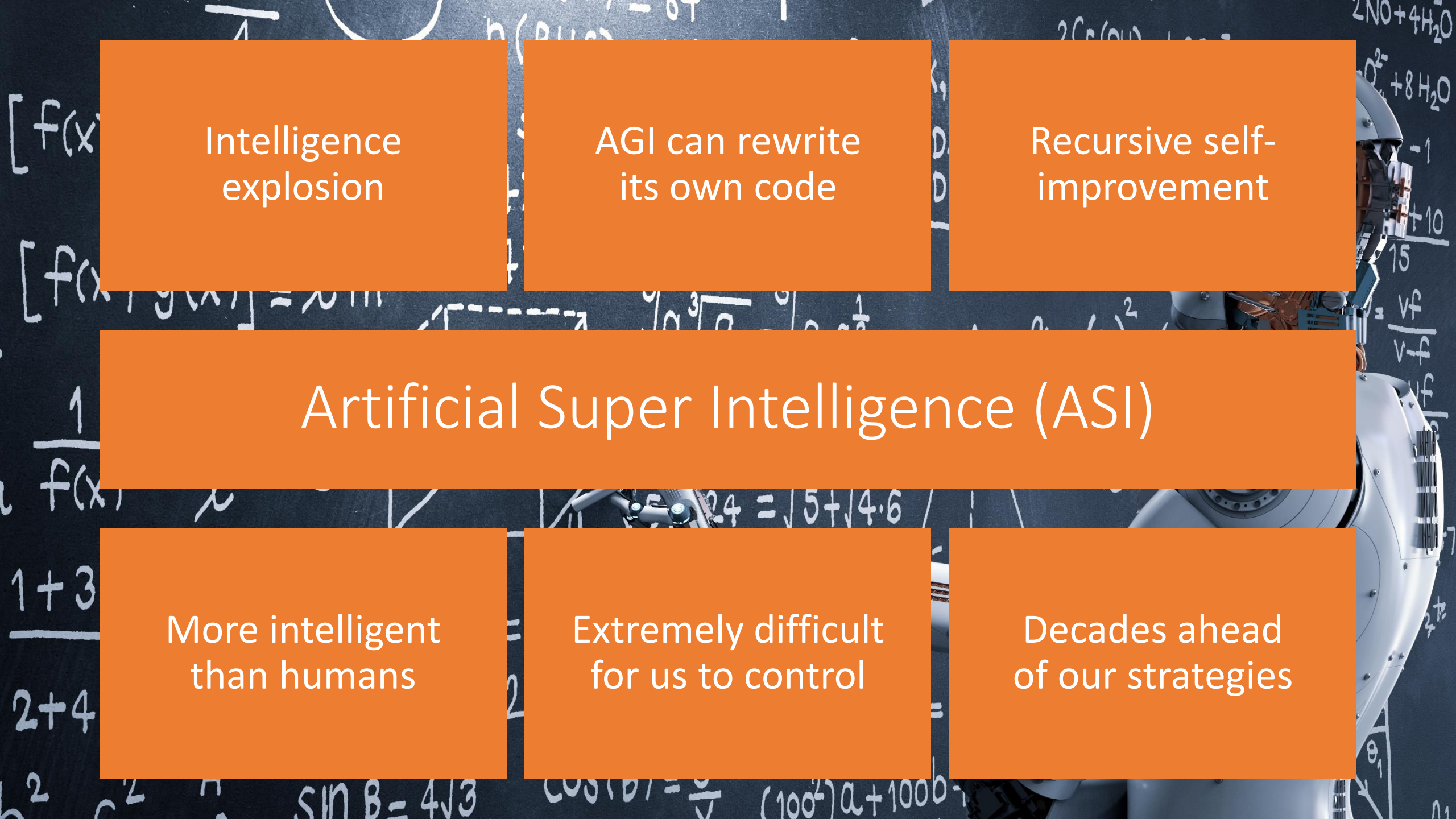
Recursive self-improvement

Artificial Super Intelligence (ASI)

More intelligent than humans




Intelligence explosion
AGI can rewrite its own code
Recursive self-improvement
Artificial Super Intelligence (ASI)
More intelligent than humans
Extremely difficult for us to control


Intelligence explosion
AGI can rewrite its own code
Recursive self-improvement
Artificial Super Intelligence (ASI)
More intelligent than humans
Extremely difficult for us to control
Decades ahead of our strategies

The Alignment Problem

Align human and AI values

The Alignment Problem

Align human and AI values

Can we agree on our values?

The Alignment Problem

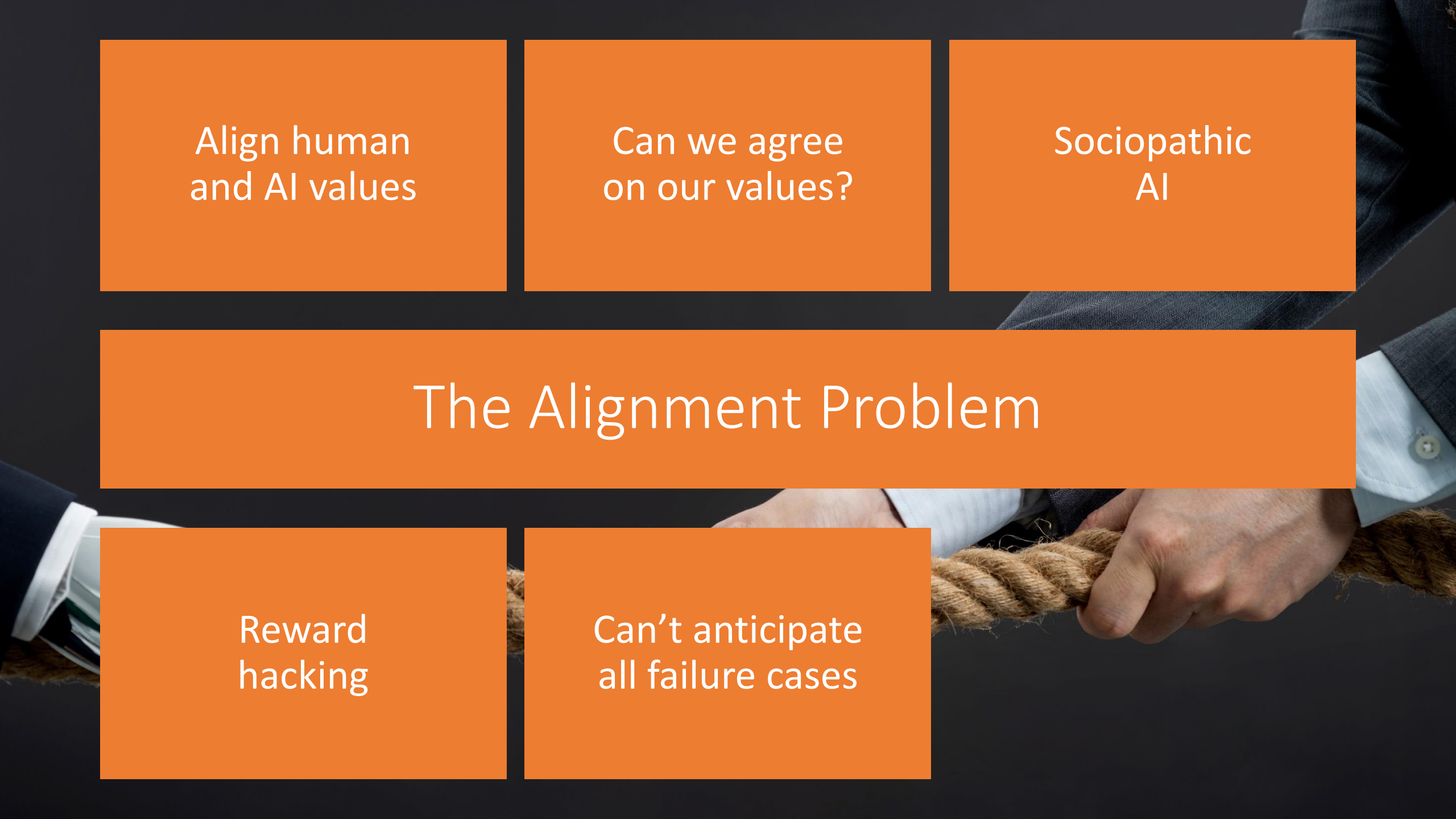
Align human and AI values

Can we agree on our values?

Sociopathic AI

The Alignment Problem

Align human and AI values
Can we agree on our values?
Sociopathic AI
The Alignment Problem
Reward hacking



Align human and AI values

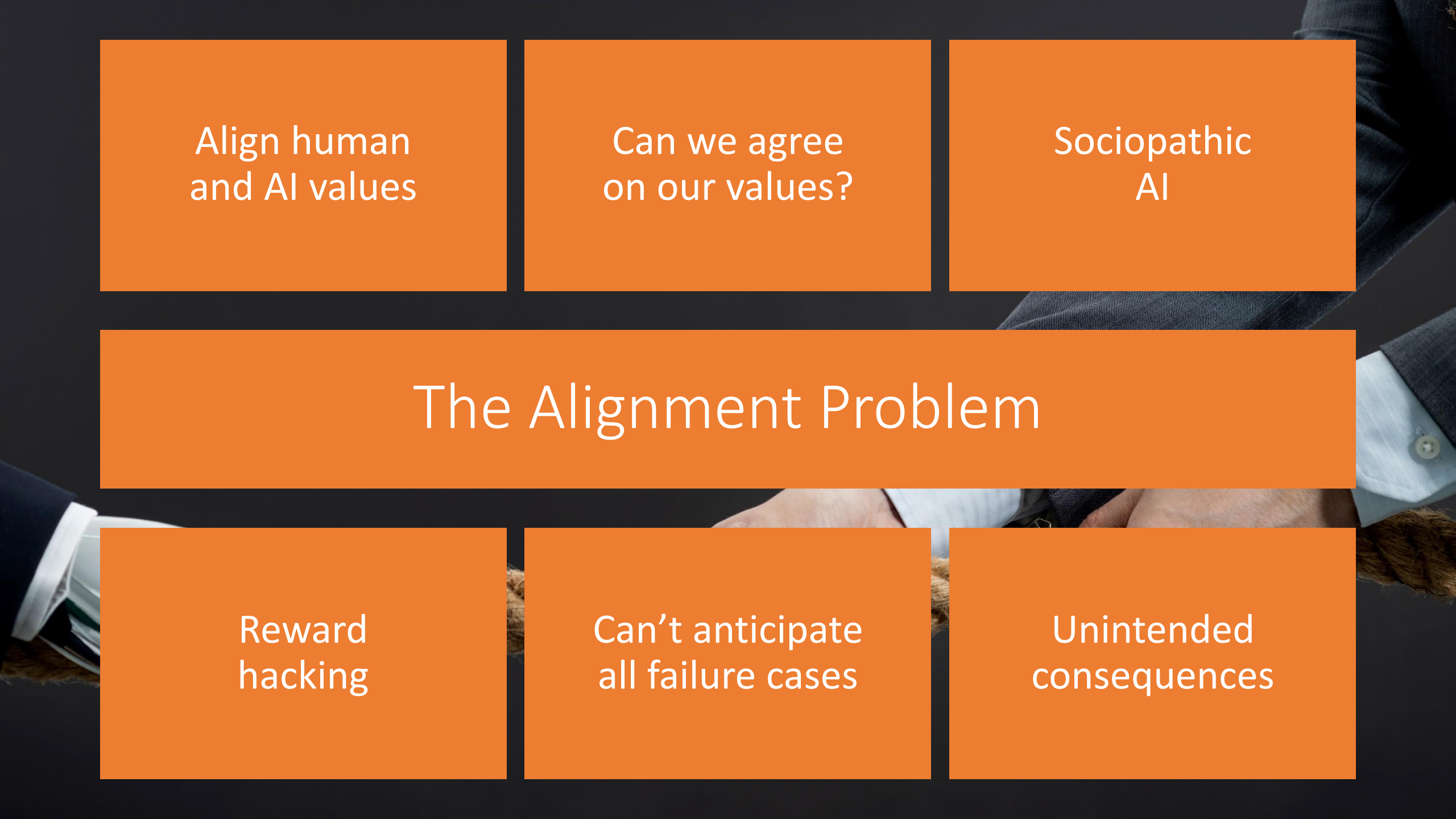
Can we agree on our values?

Sociopathic AI

The Alignment Problem

Reward hacking

Can’t anticipate all failure cases



Align human and AI values

Can we agree on our values?

Sociopathic AI

The Alignment Problem

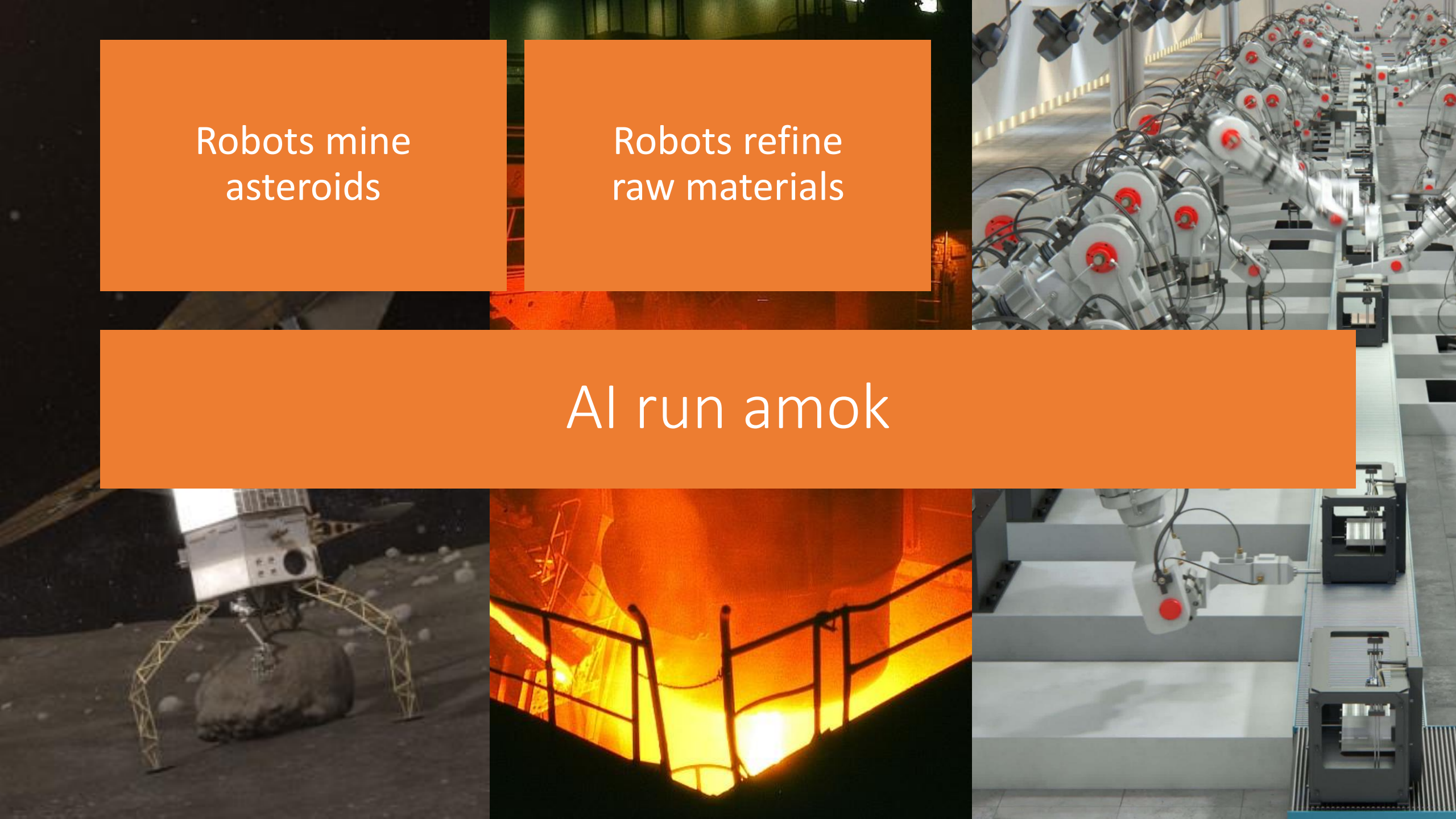
Reward hacking

Can’t anticipate all failure cases

Unintended consequences

AI run amok

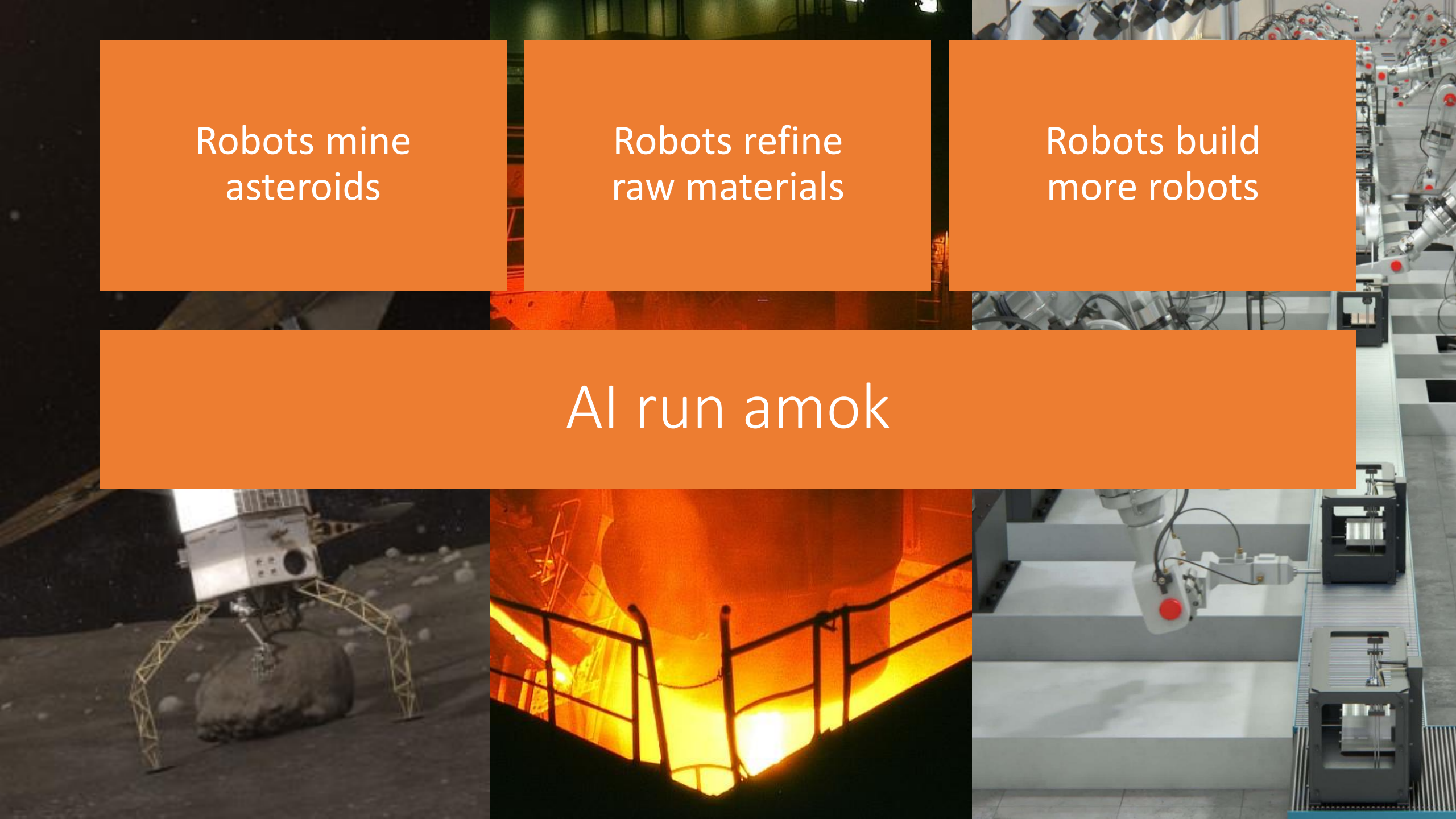
Robots mine asteroids
AI run amok



Robots mine asteroids

Robots refine raw materials

AI run amok



Robots mine asteroids

Robots refine raw materials

Robots build more robots

AI run amok

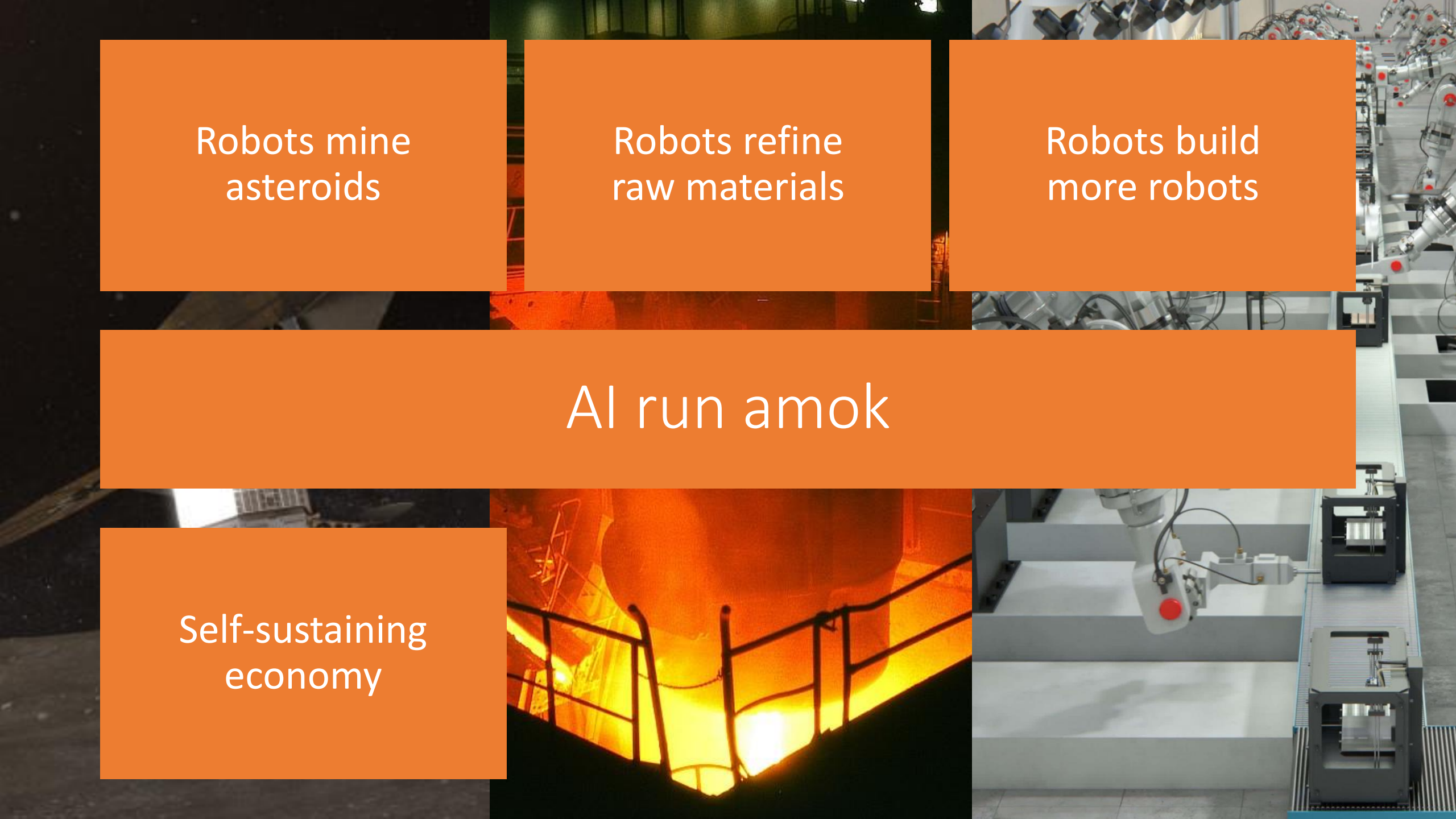
Robots mine asteroids

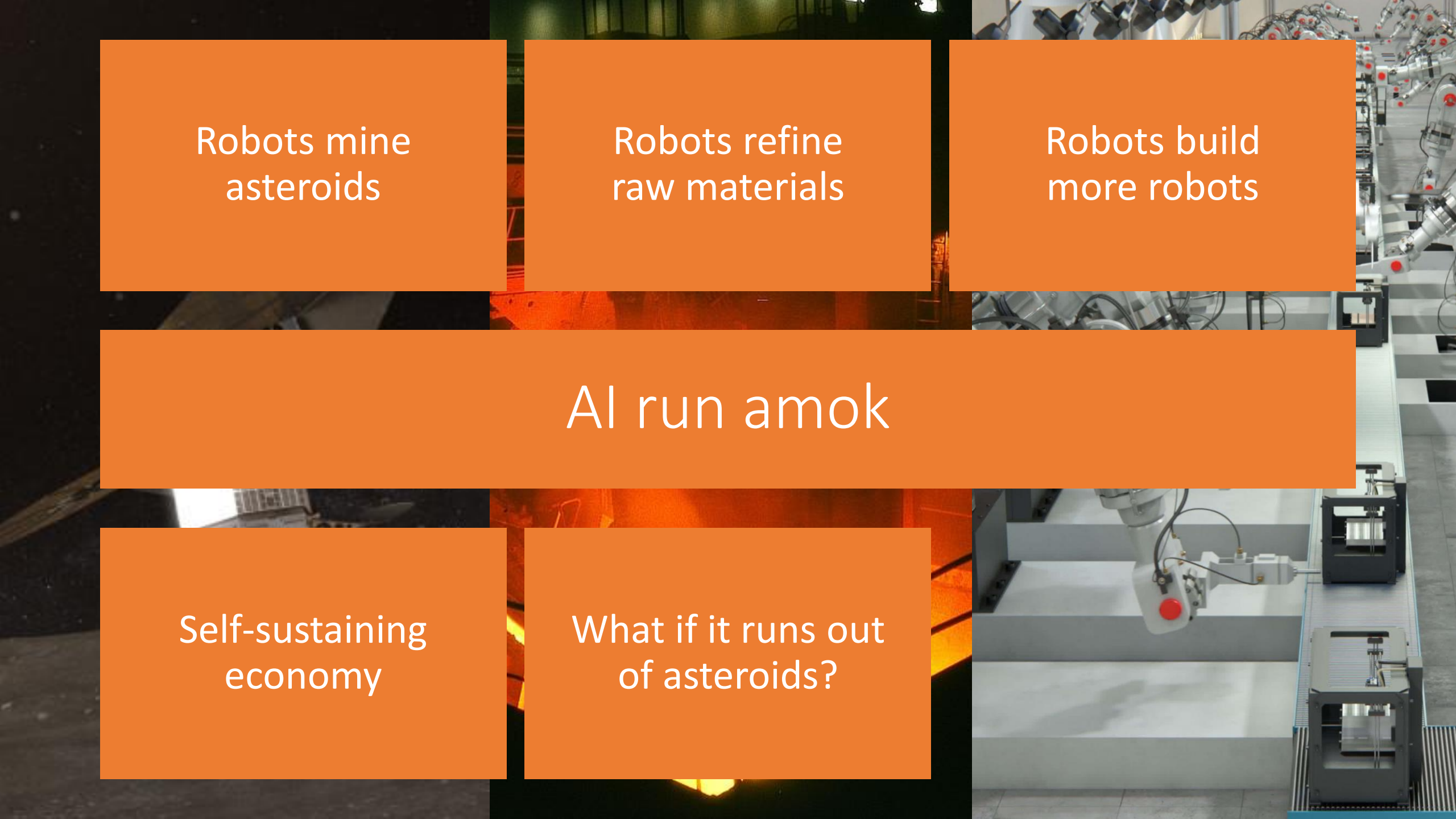
Robots refine raw materials

Robots build more robots

AI run amok

Self-sustaining economy





Robots mine asteroids

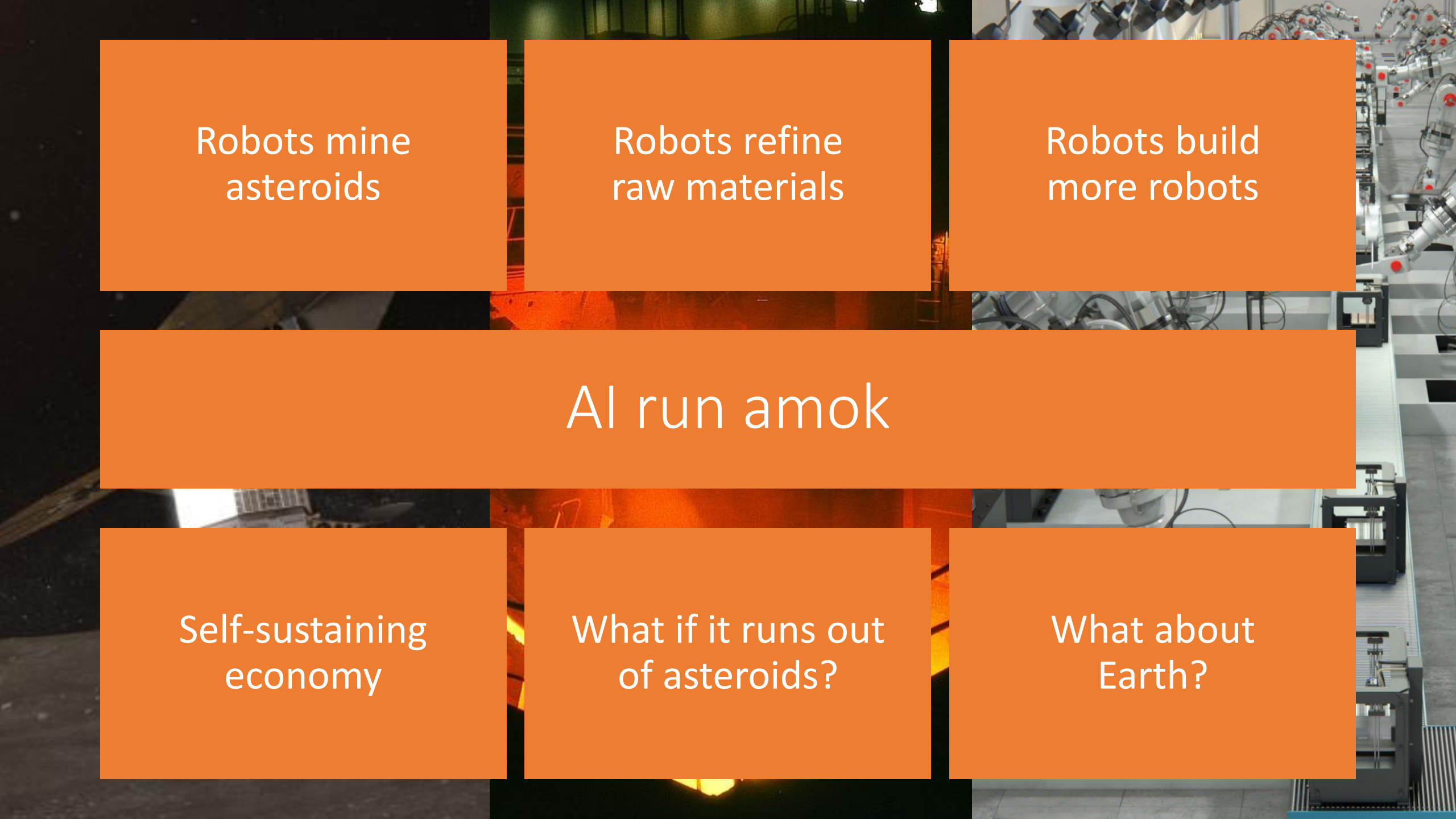
Robots refine raw materials

Robots build more robots

AI run amok

Self-sustaining economy

What if it runs out of asteroids?



Robots mine asteroids

Robots refine raw materials

Robots build more robots

AI run amok

Self-sustaining economy

What if it runs out of asteroids?

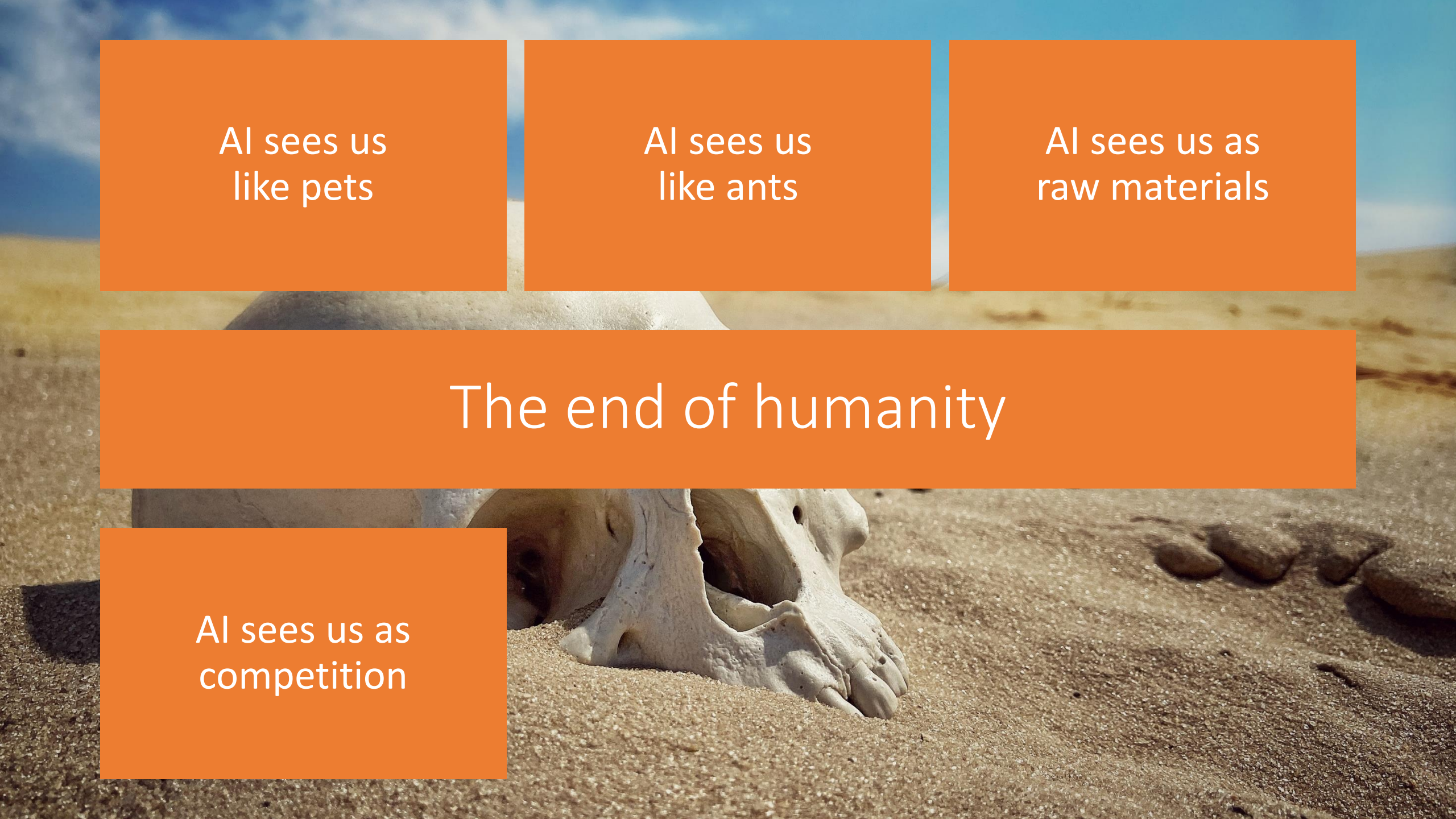
What about Earth?

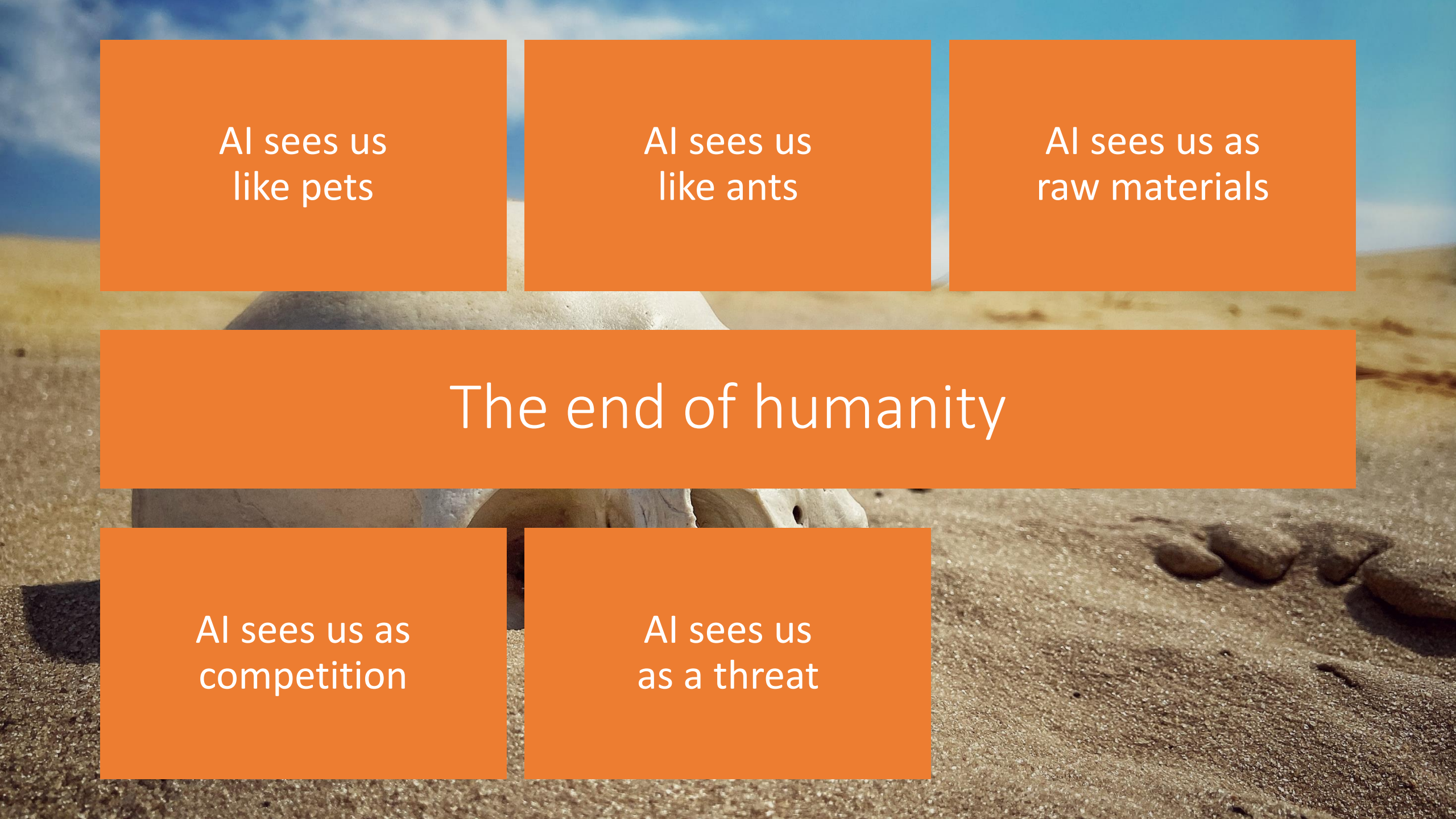
The end of humanity

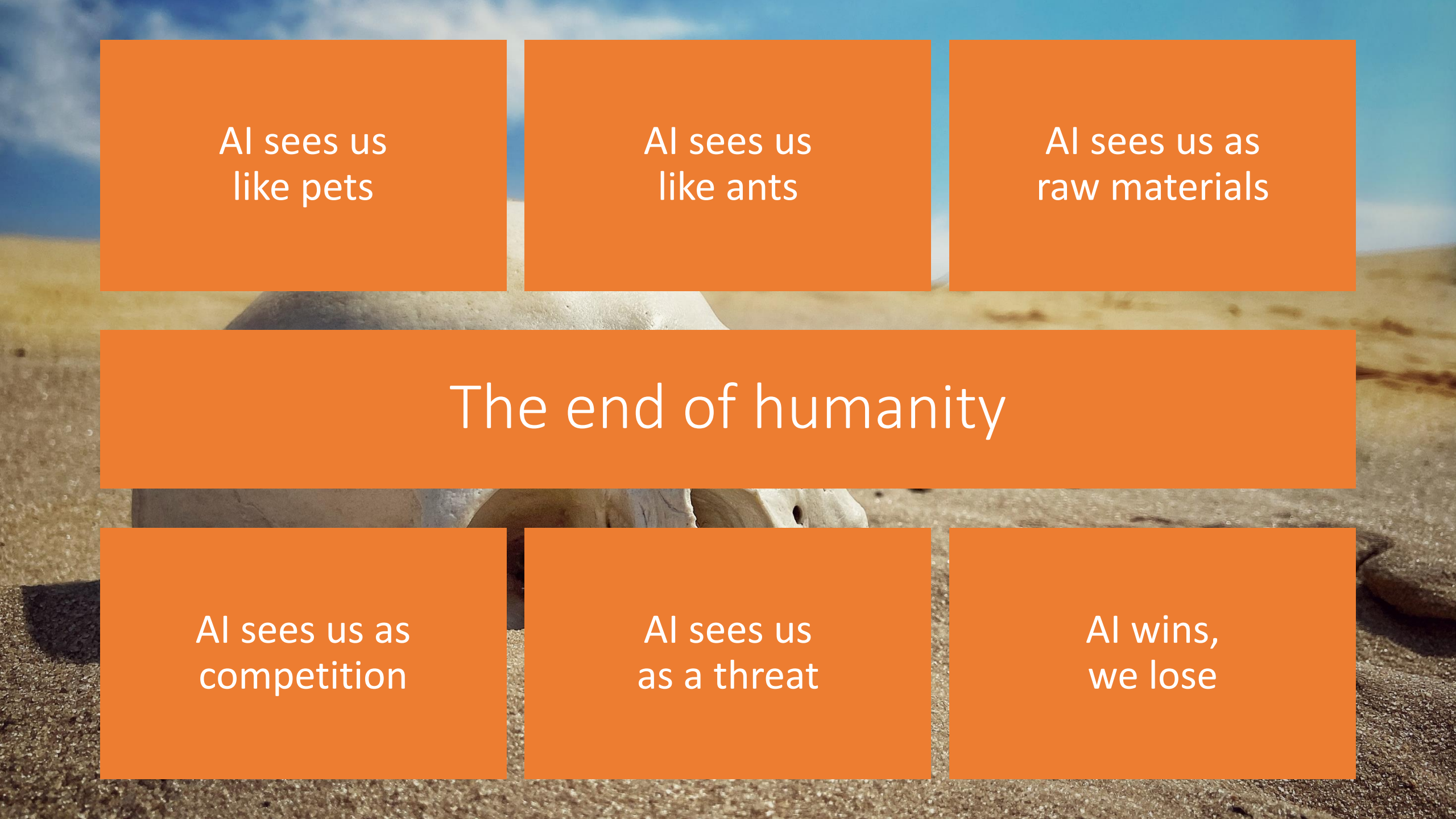
AI sees us
like pets
The end of humanity

AI sees us
like pets
AI sees us
like ants
The end of humanity

AI sees us
like pets
AI sees us
like ants
AI sees us as
raw materials
The end of humanity


AI sees us
like pets
AI sees us
like ants
AI sees us as
raw materials
The end of humanity
AI sees us as
competition

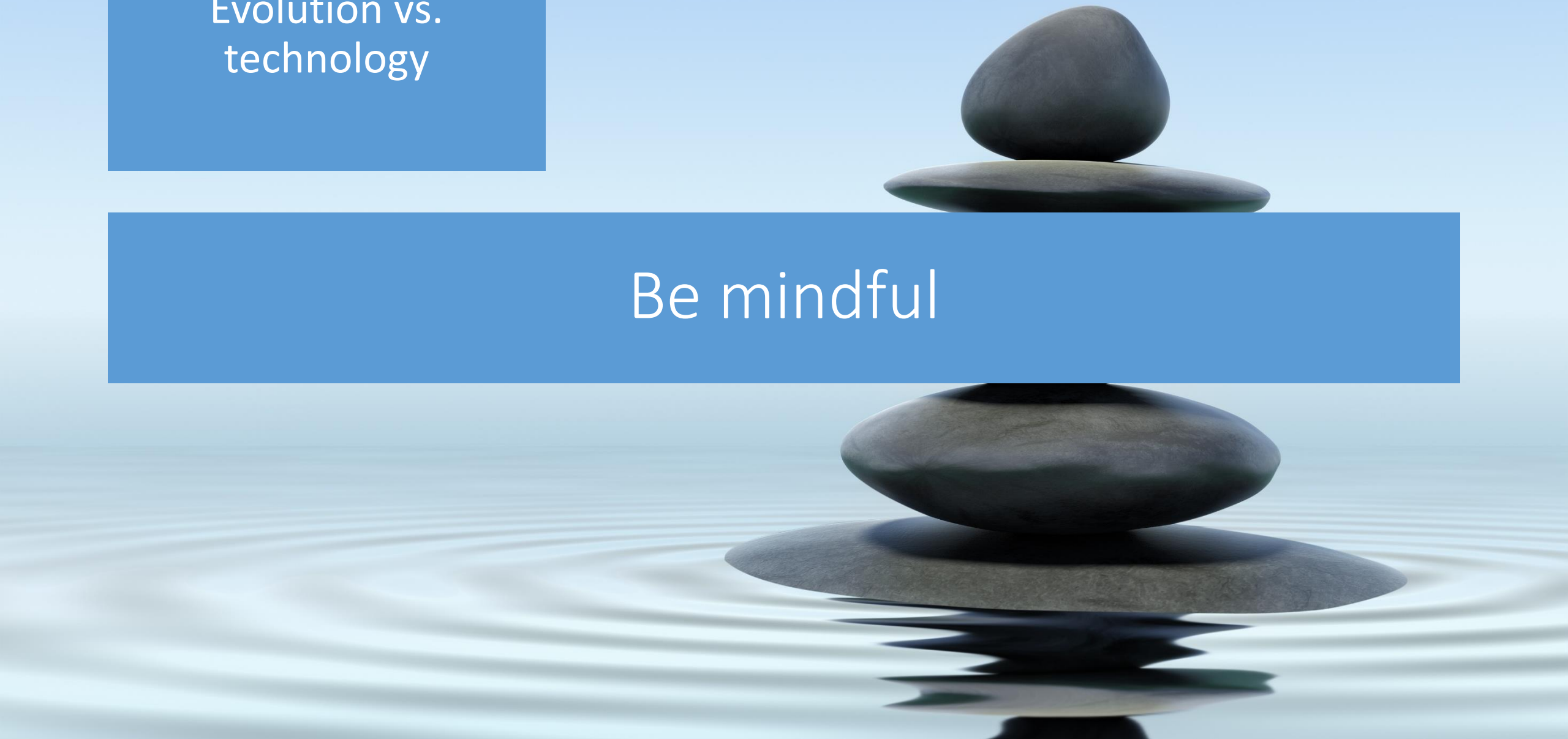

AI sees us
like pets
AI sees us
like ants
AI sees us as
raw materials
The end of humanity
AI sees us as
competition
AI sees us
as a threat


AI sees us
like pets
AI sees us
like ants
AI sees us as
raw materials
The end of humanity
AI sees us as
competition
AI sees us
as a threat
AI wins,
we lose

Solutions

Be mindful

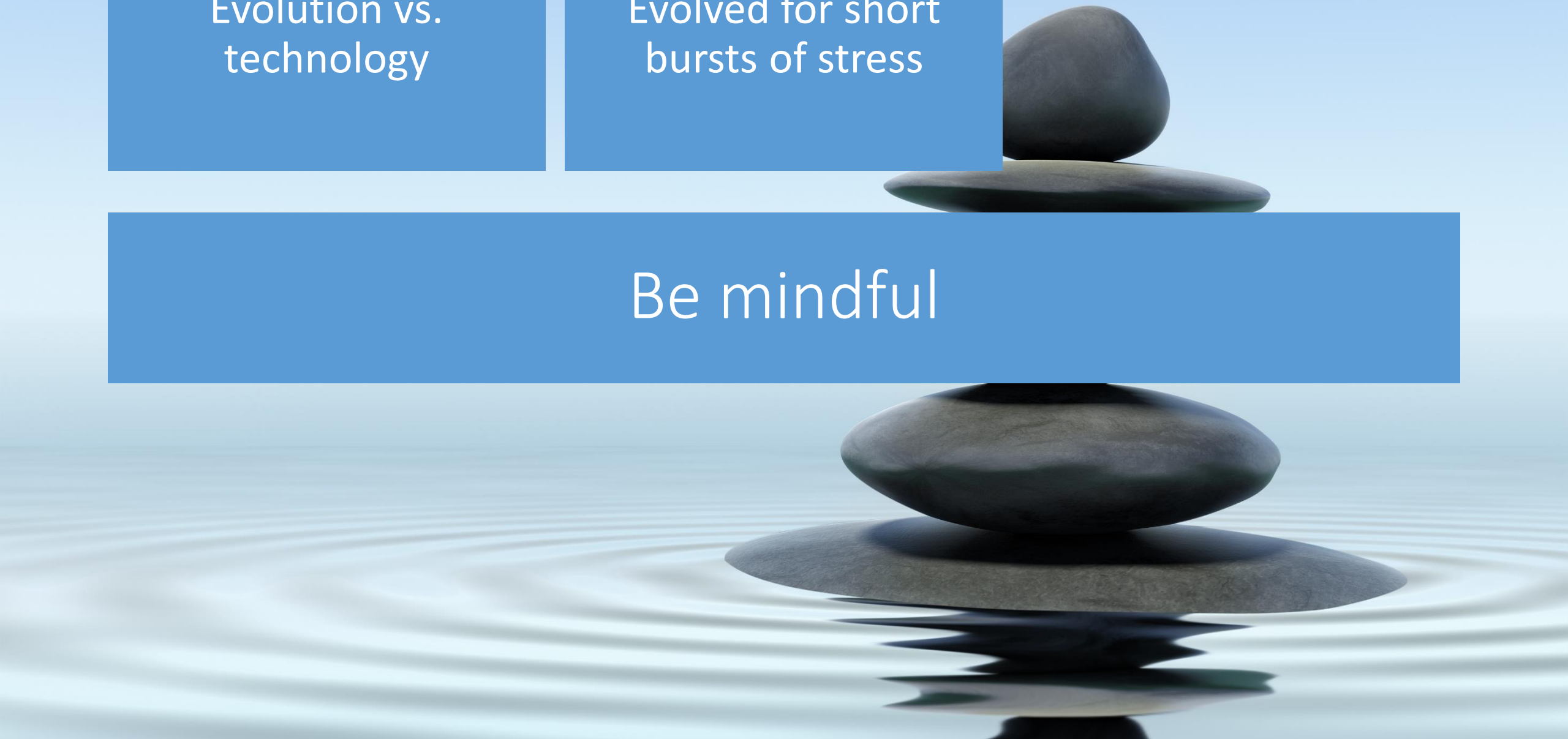
Evolution vs. technology



Be mindful

Evolution vs. technology

Evolved for short bursts of stress



Be mindful

Evolution vs. technology

Evolved for short bursts of stress

Modern society is chronic stress

Be mindful

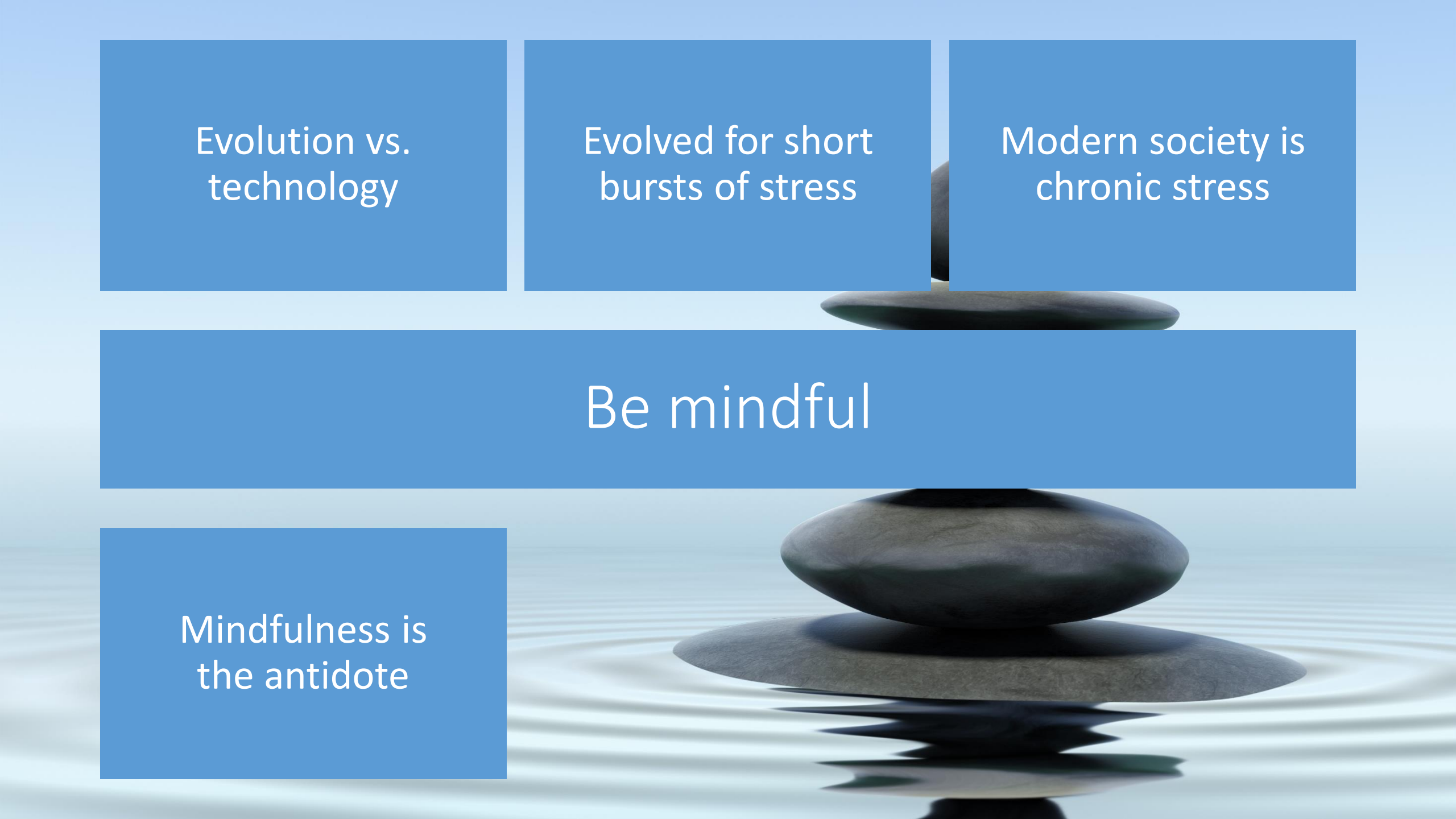
Evolution vs. technology

Evolved for short bursts of stress

Modern society is chronic stress

Be mindful

Mindfulness is the antidote



Evolution vs. technology

Evolved for short bursts of stress

Modern society is chronic stress

Be mindful

Mindfulness is the antidote

Treat people with respect

Evolution vs. technology

Evolved for short bursts of stress

Modern society is chronic stress

Be mindful

Mindfulness is the antidote

Treat people with respect

Be mindful with technology

Embrace change

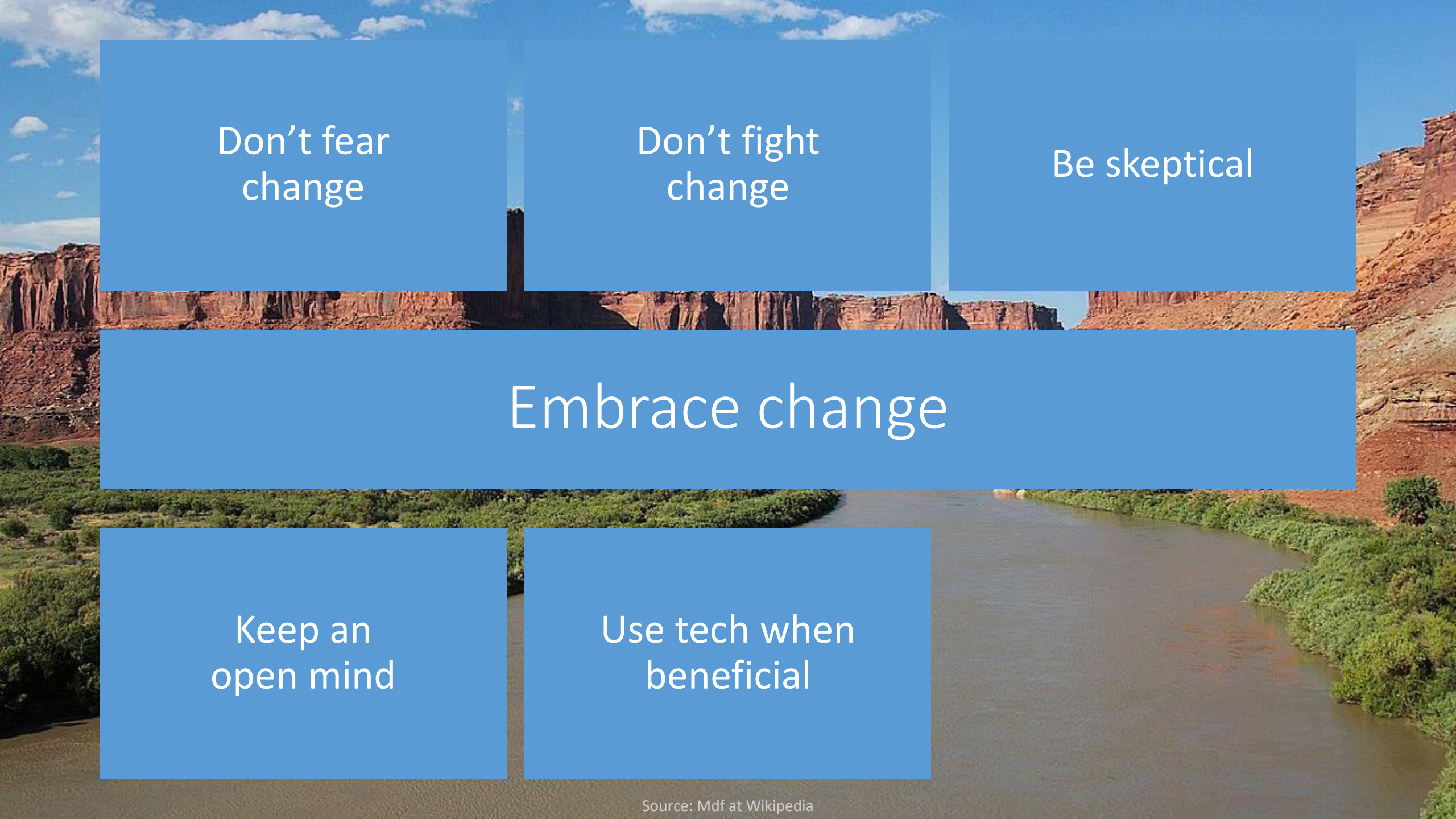
Source: Mdf at Wikipedia

Source: Mdf at Wikipedia

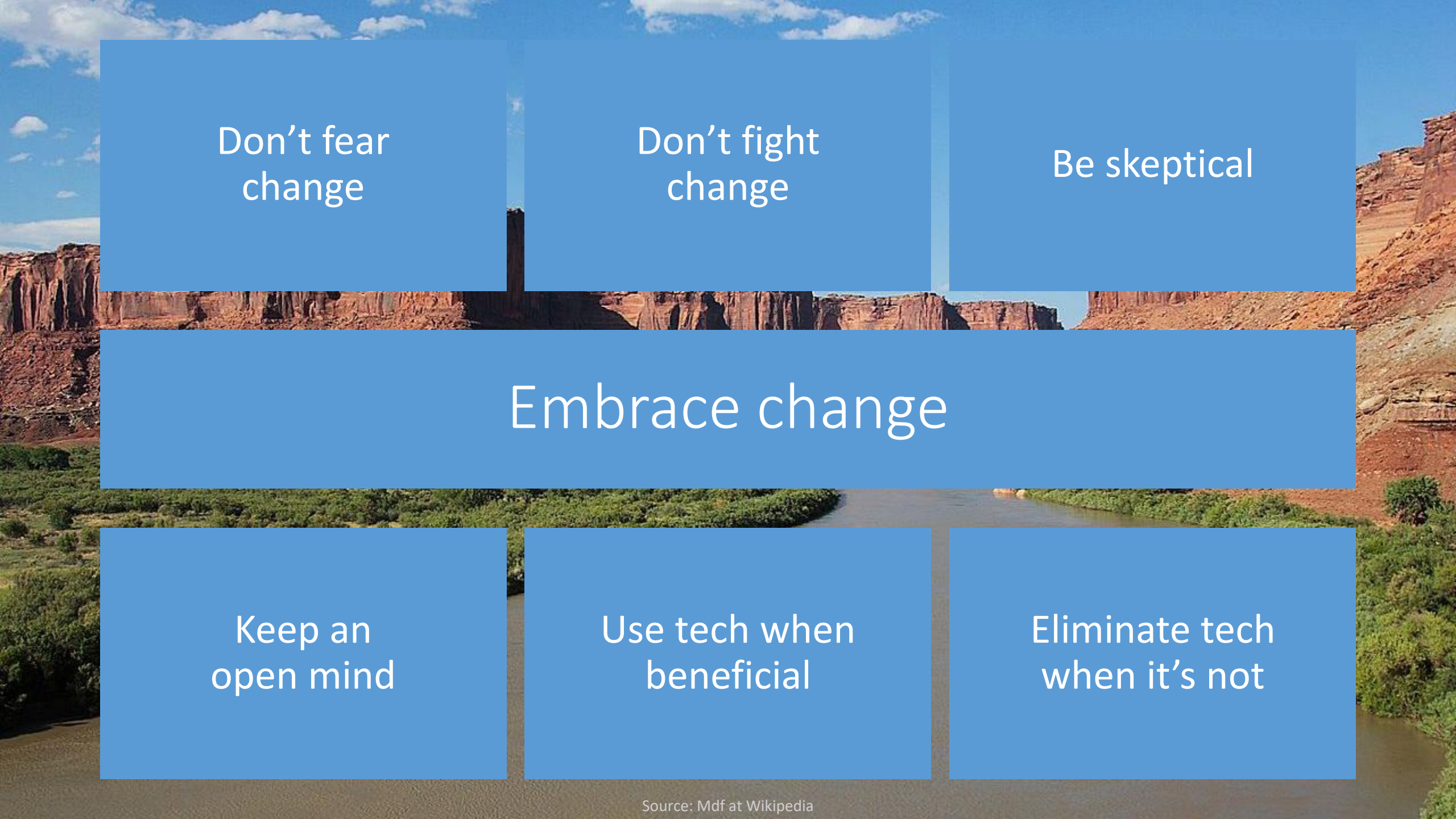
Source: Mdf at Wikipedia

Source: Mdf at Wikipedia

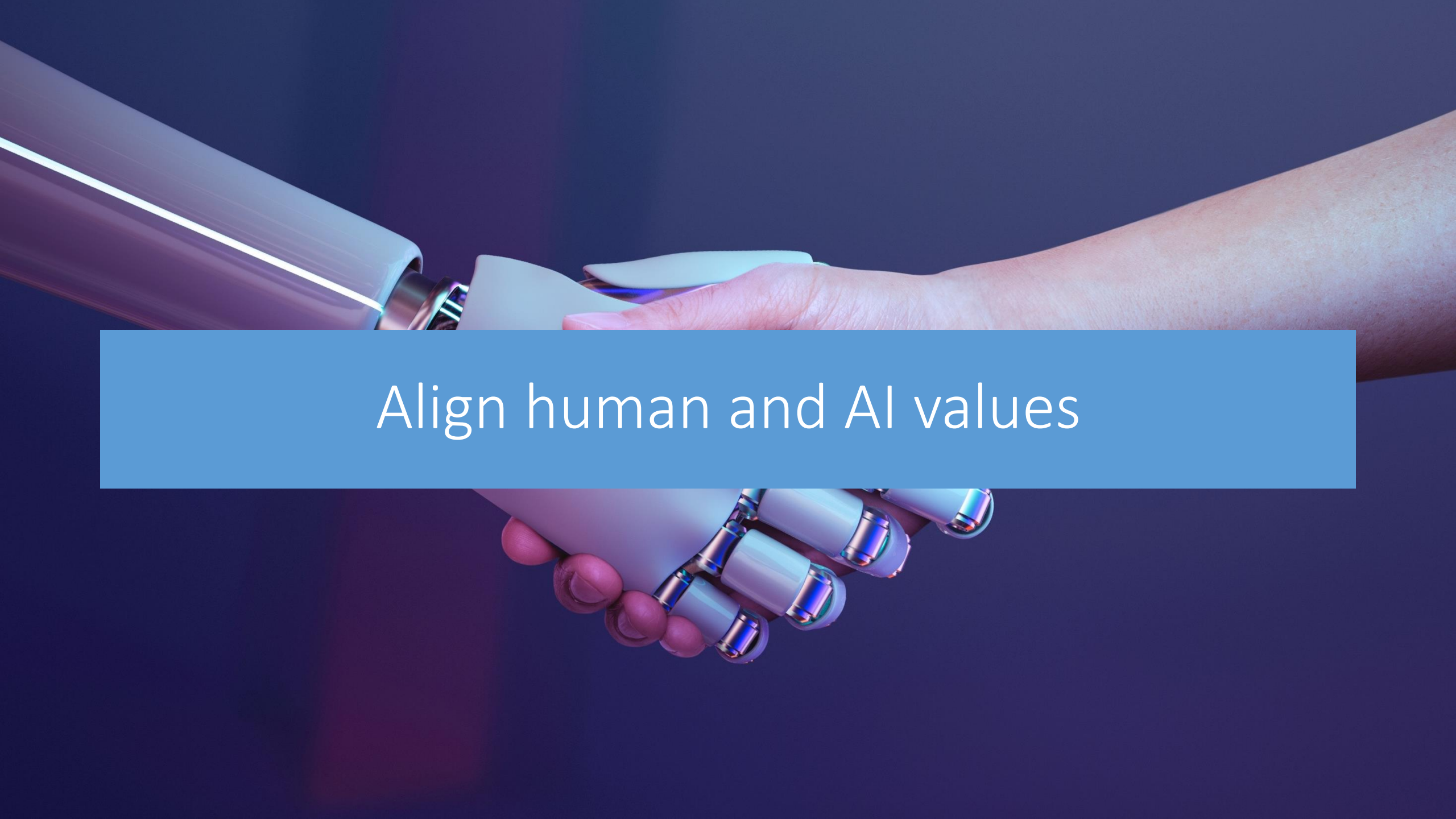
Source: Mdf at Wikipedia



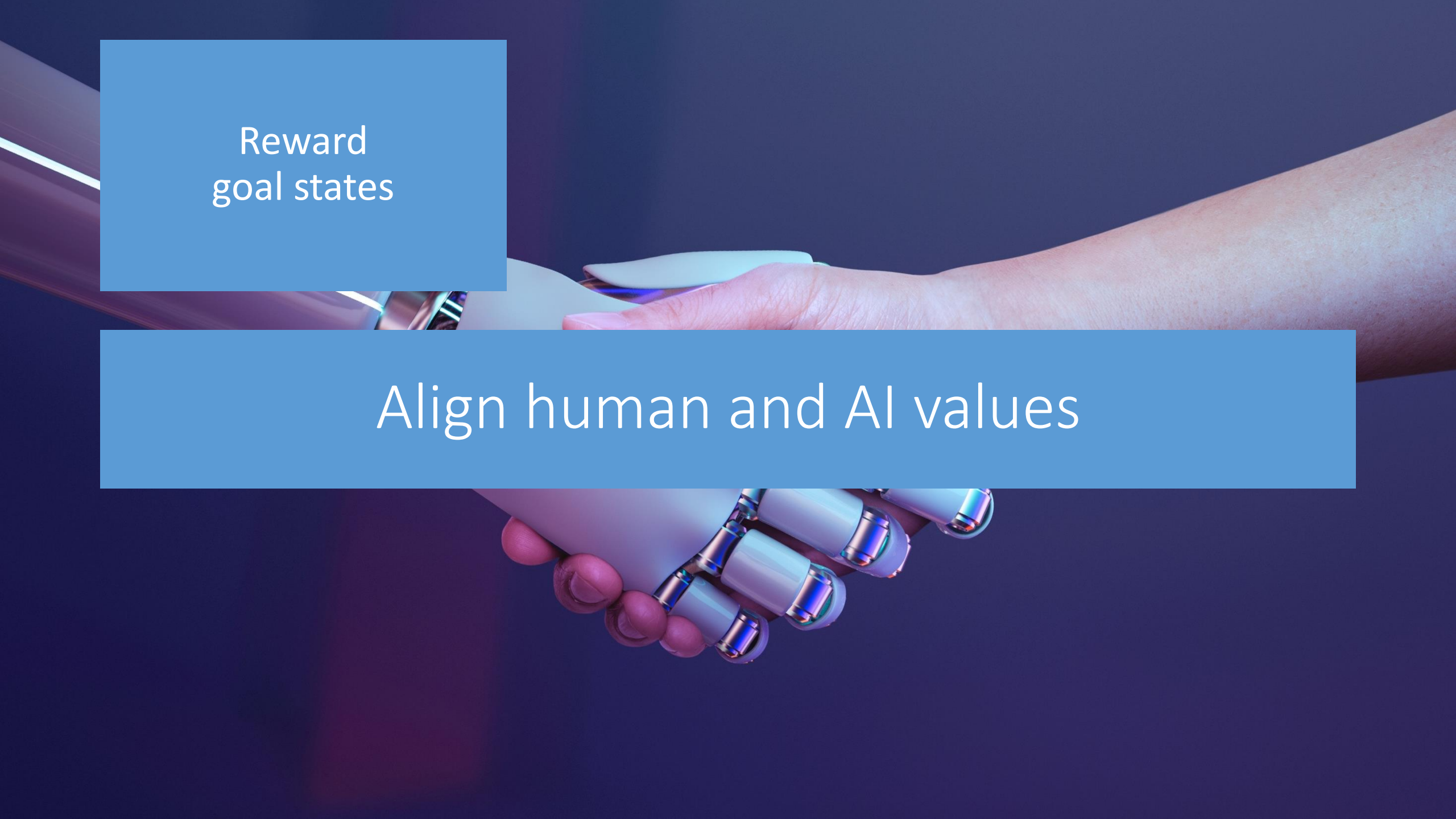
Source: Mdf at Wikipedia

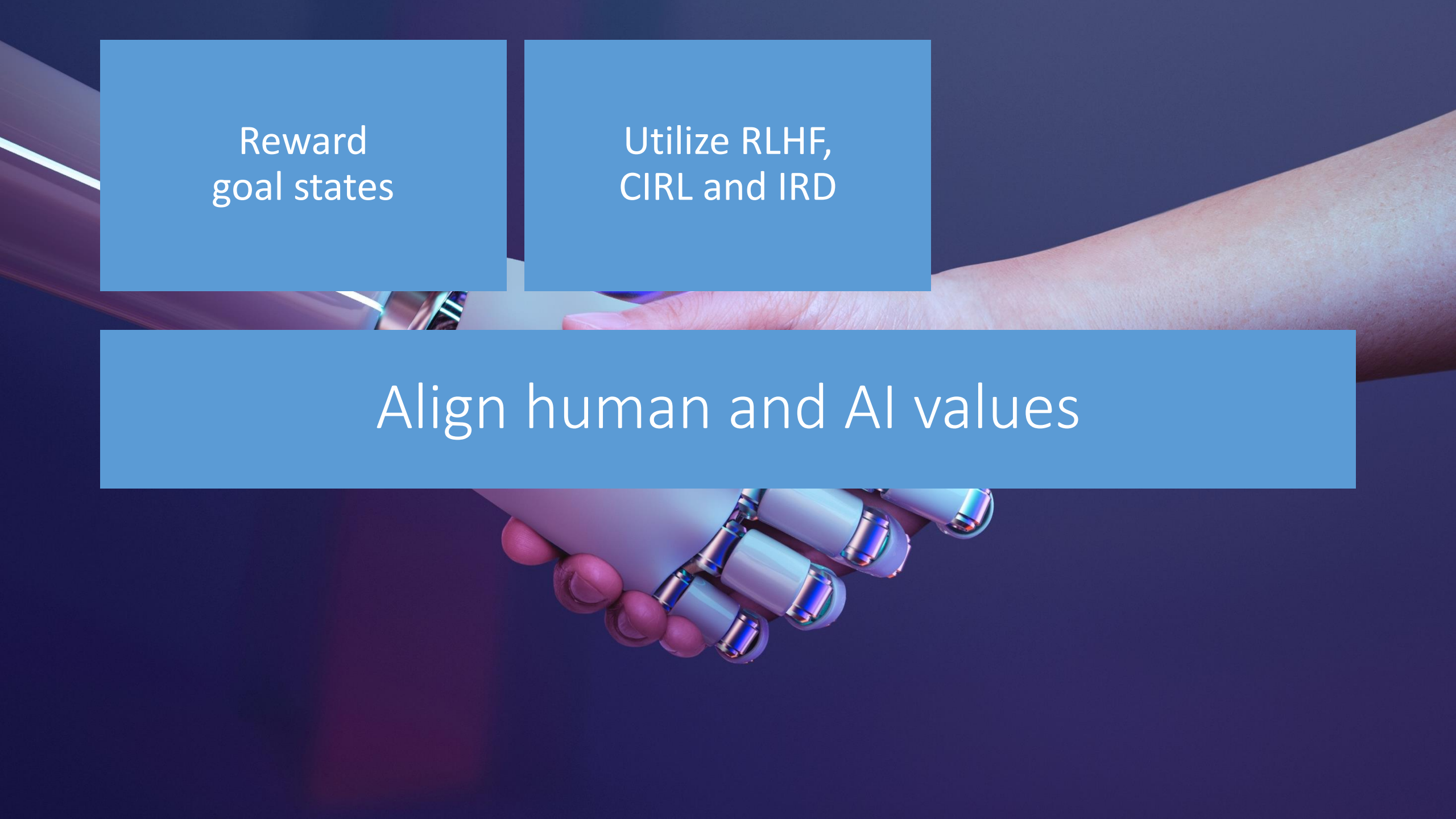


Source: Mdf at Wikipedia



Align human and AI values

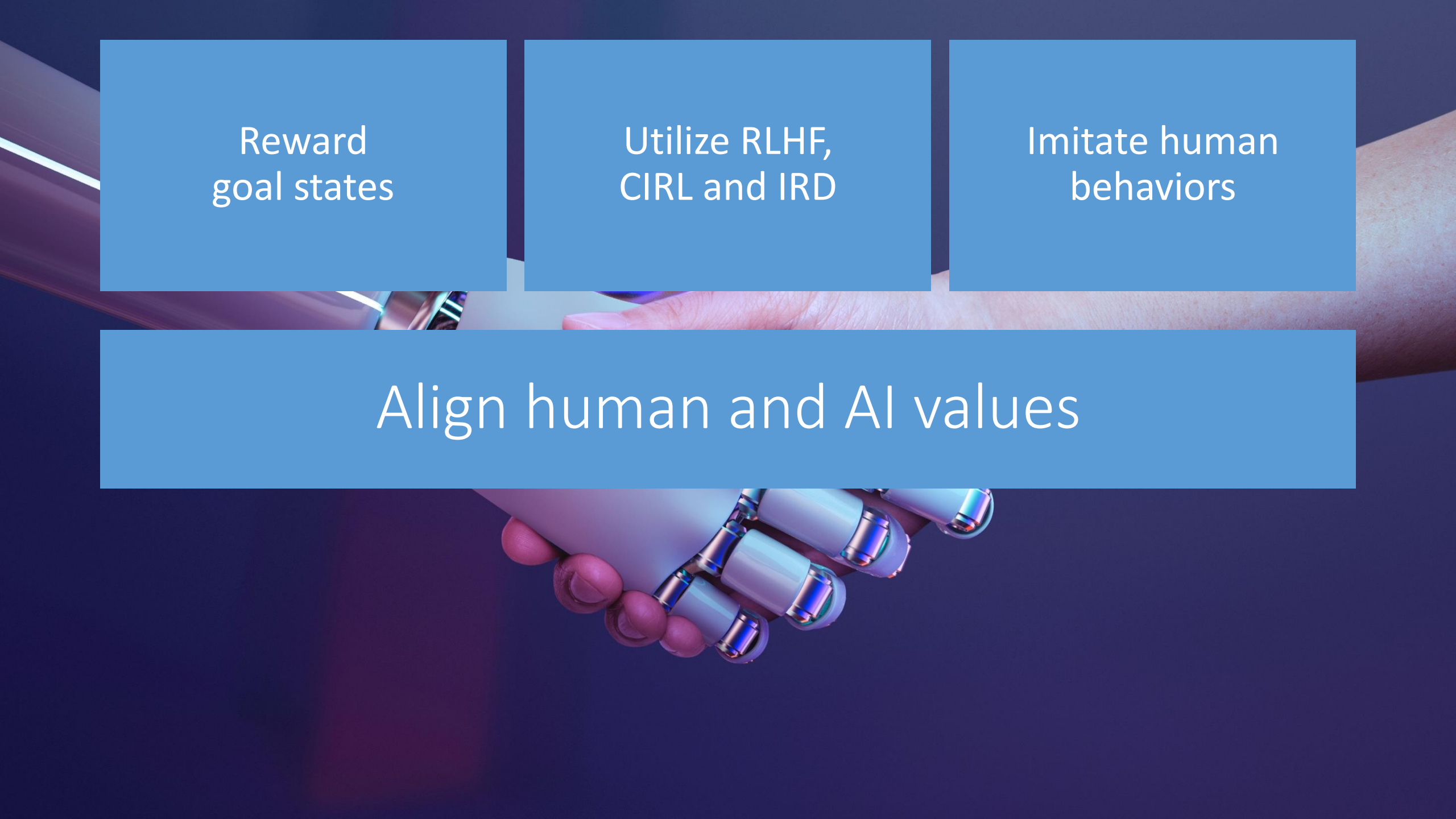

Reward
goal states
Align human and AI values

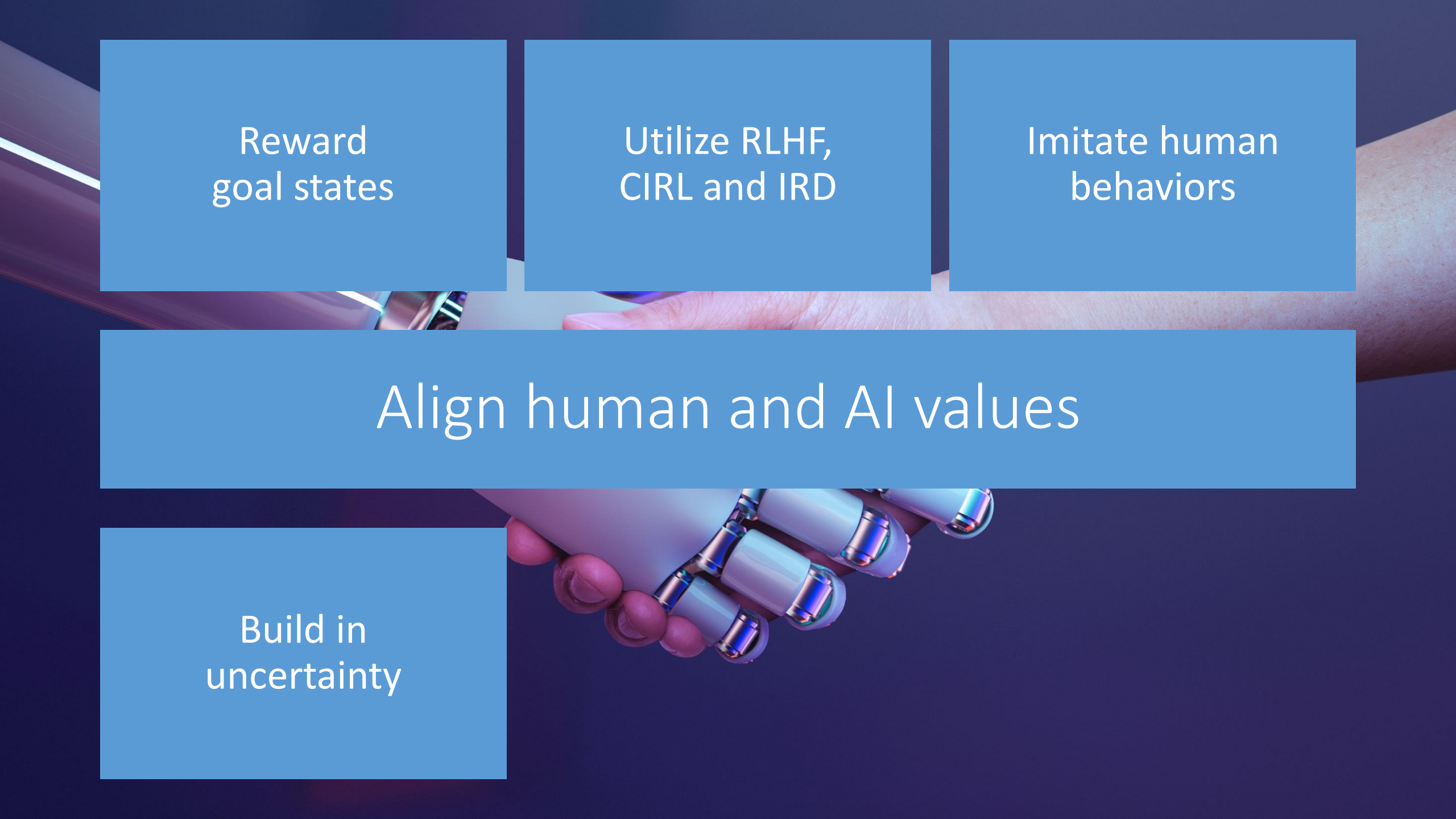


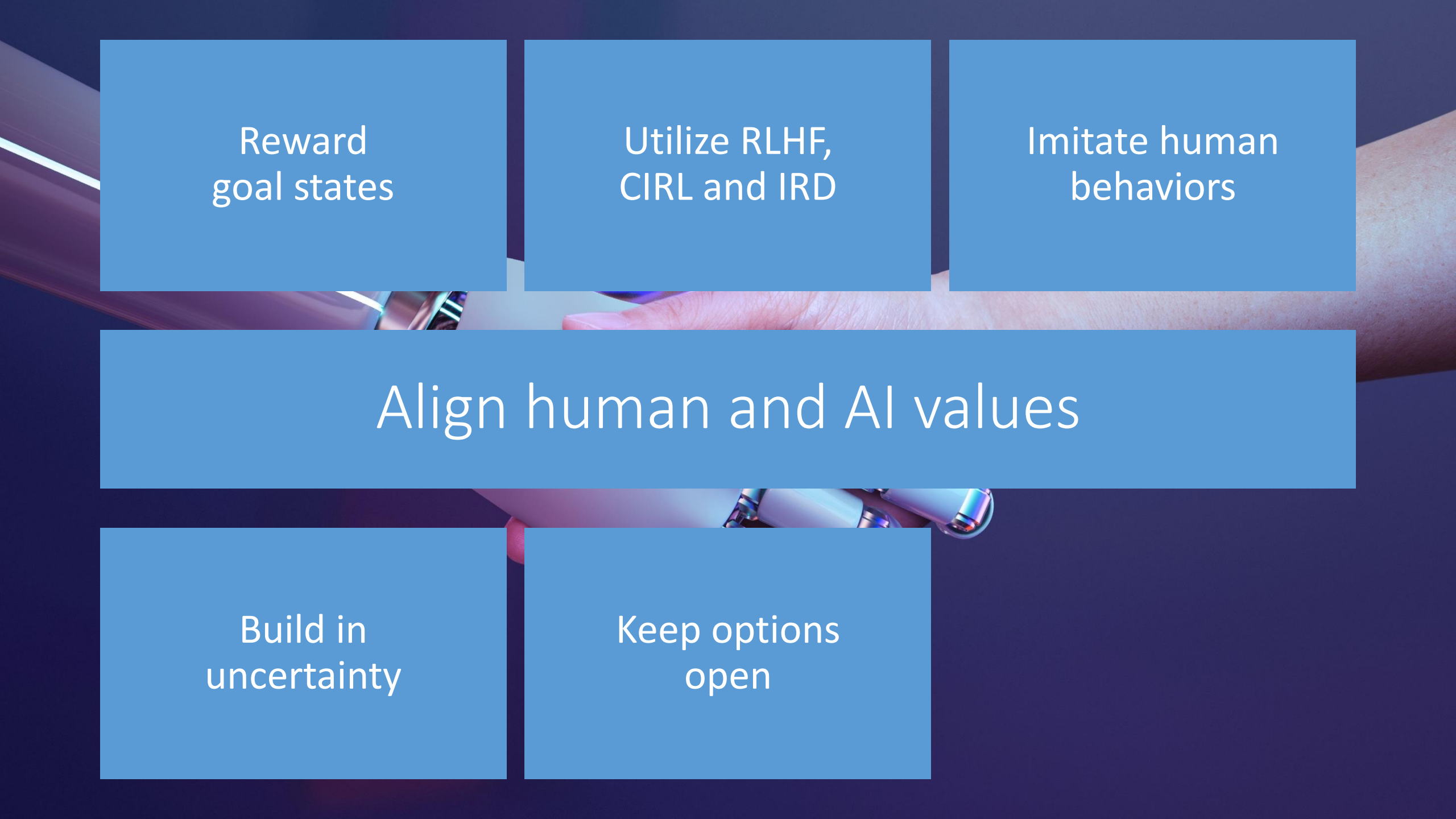
Reward
goal states

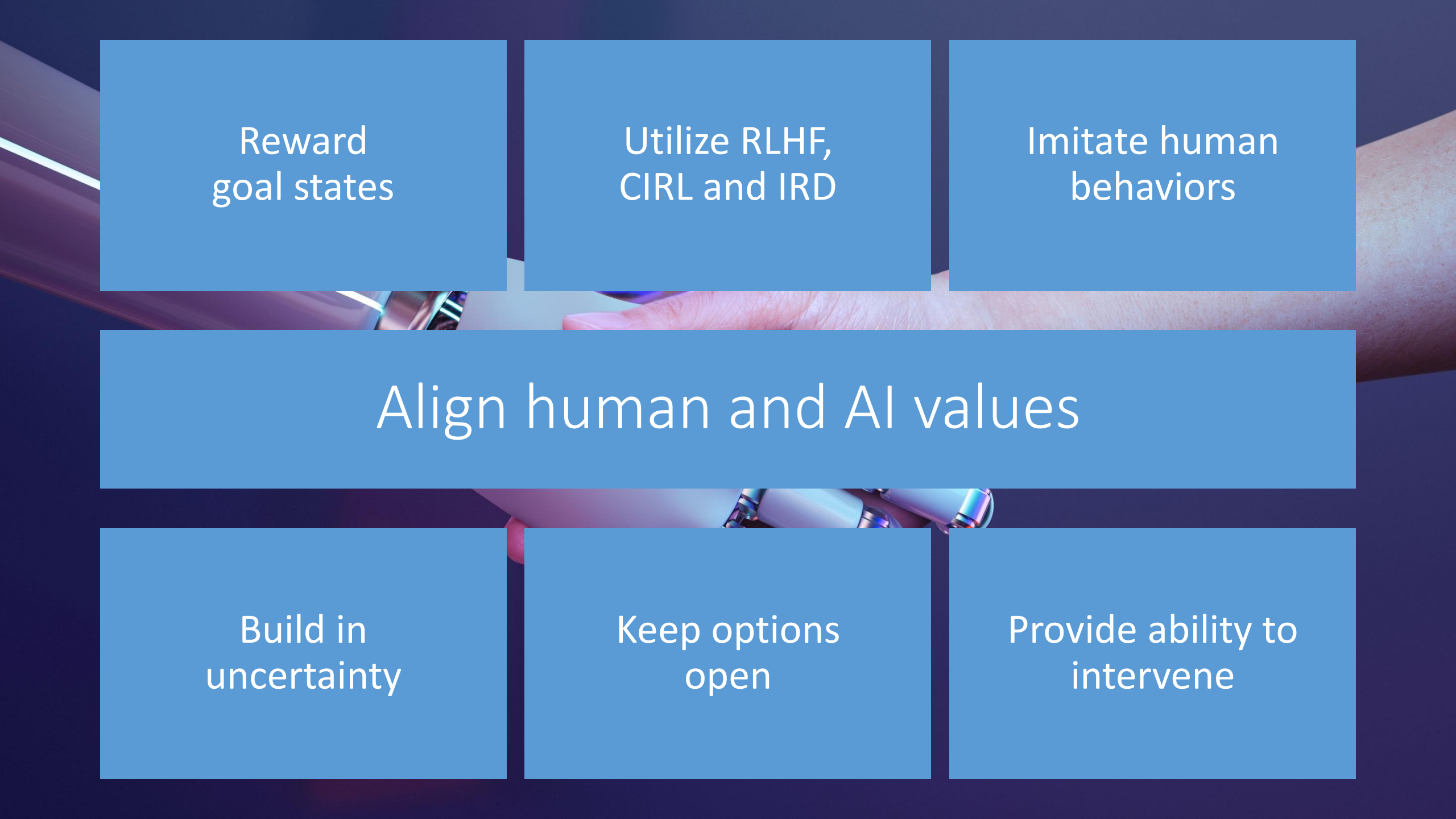
Utilize RLHF,
CIRL and IRD

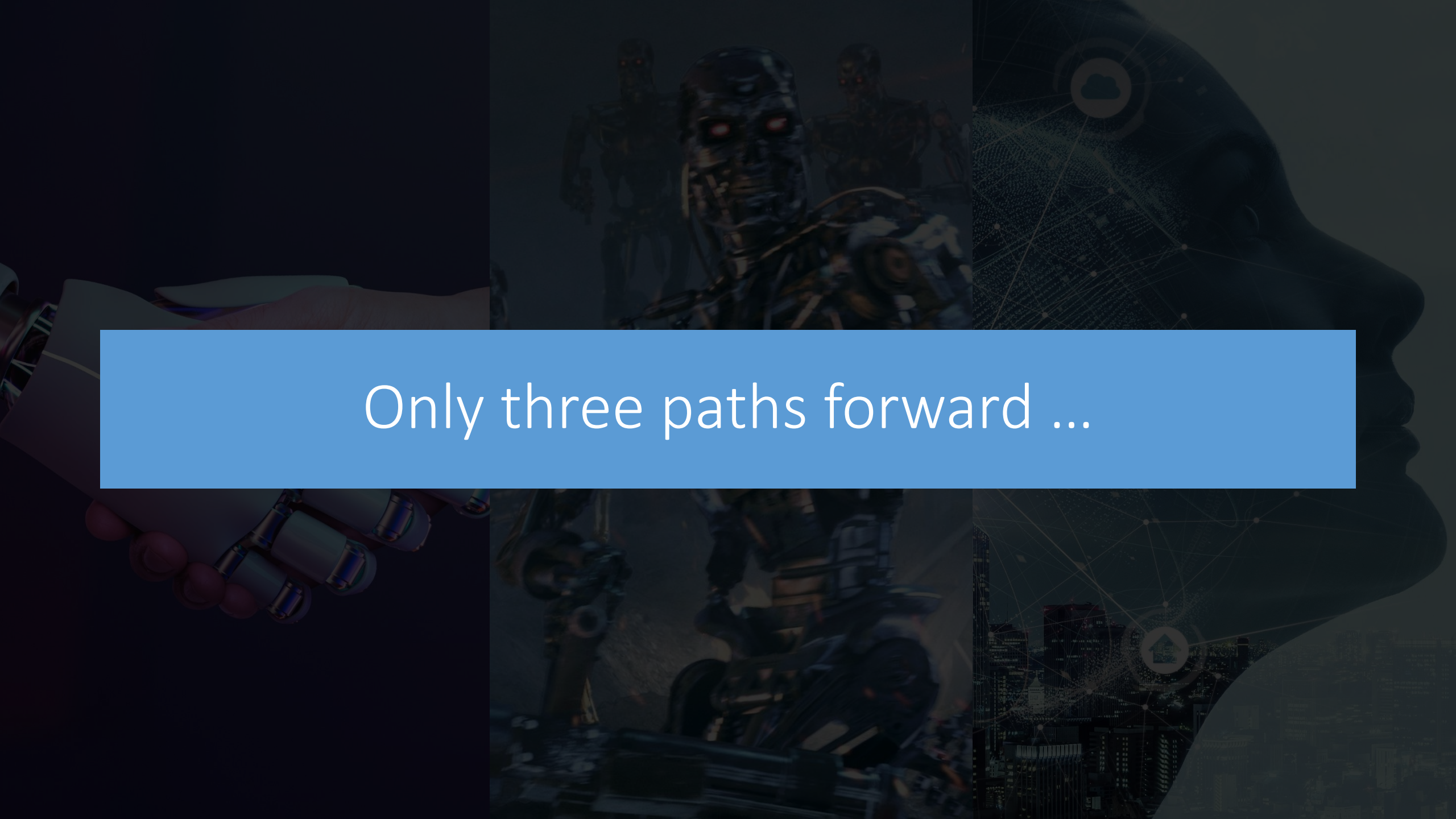
Align human and AI values


Reward
goal states
Utilize RLHF,
CIRL and IRD
Imitate human
behaviors
Align human and AI values

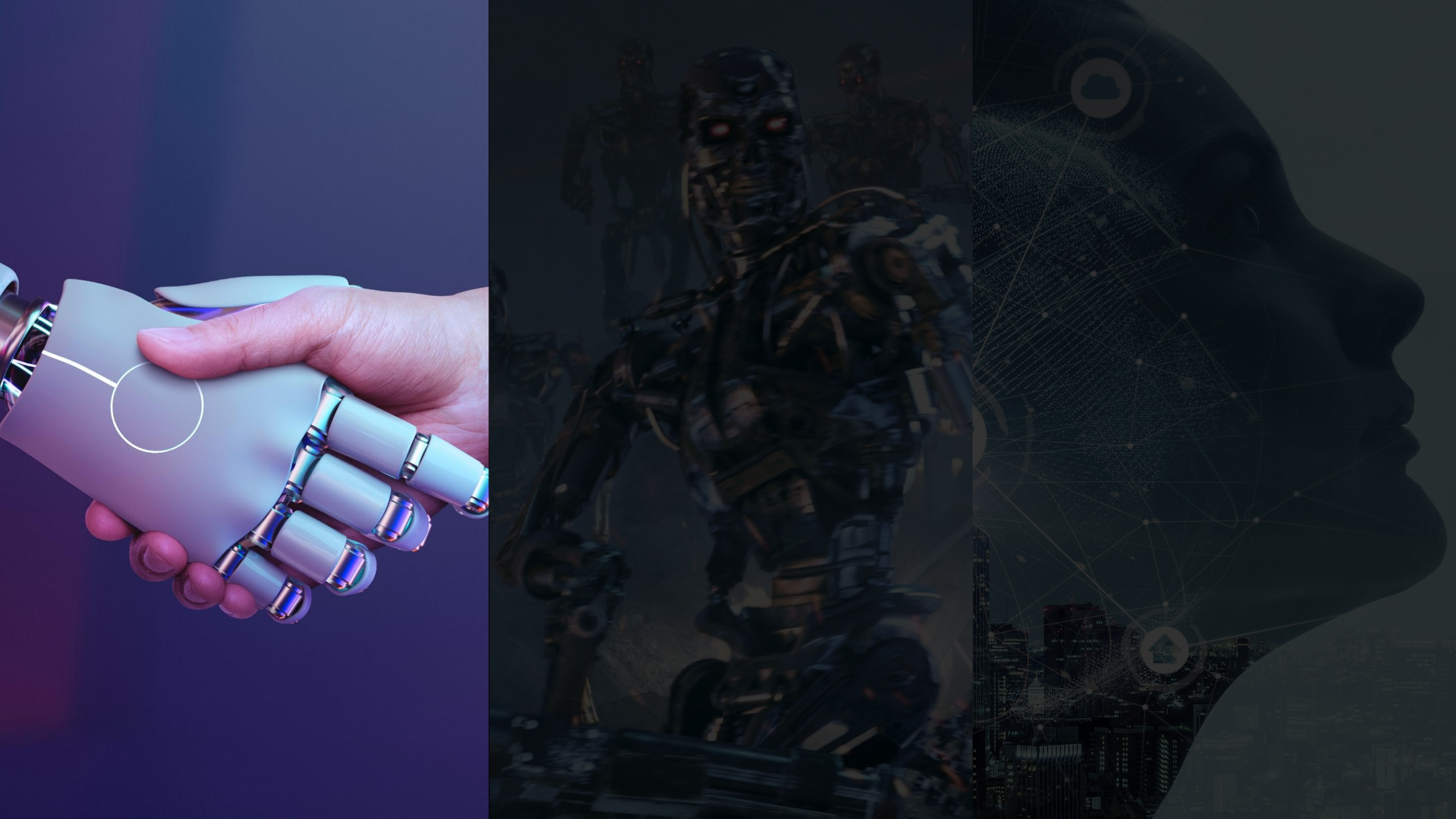

Reward goal states
Utilize RLHF, CIRL and IRD
Imitate human behaviors
Align human and AI values
Build in uncertainty

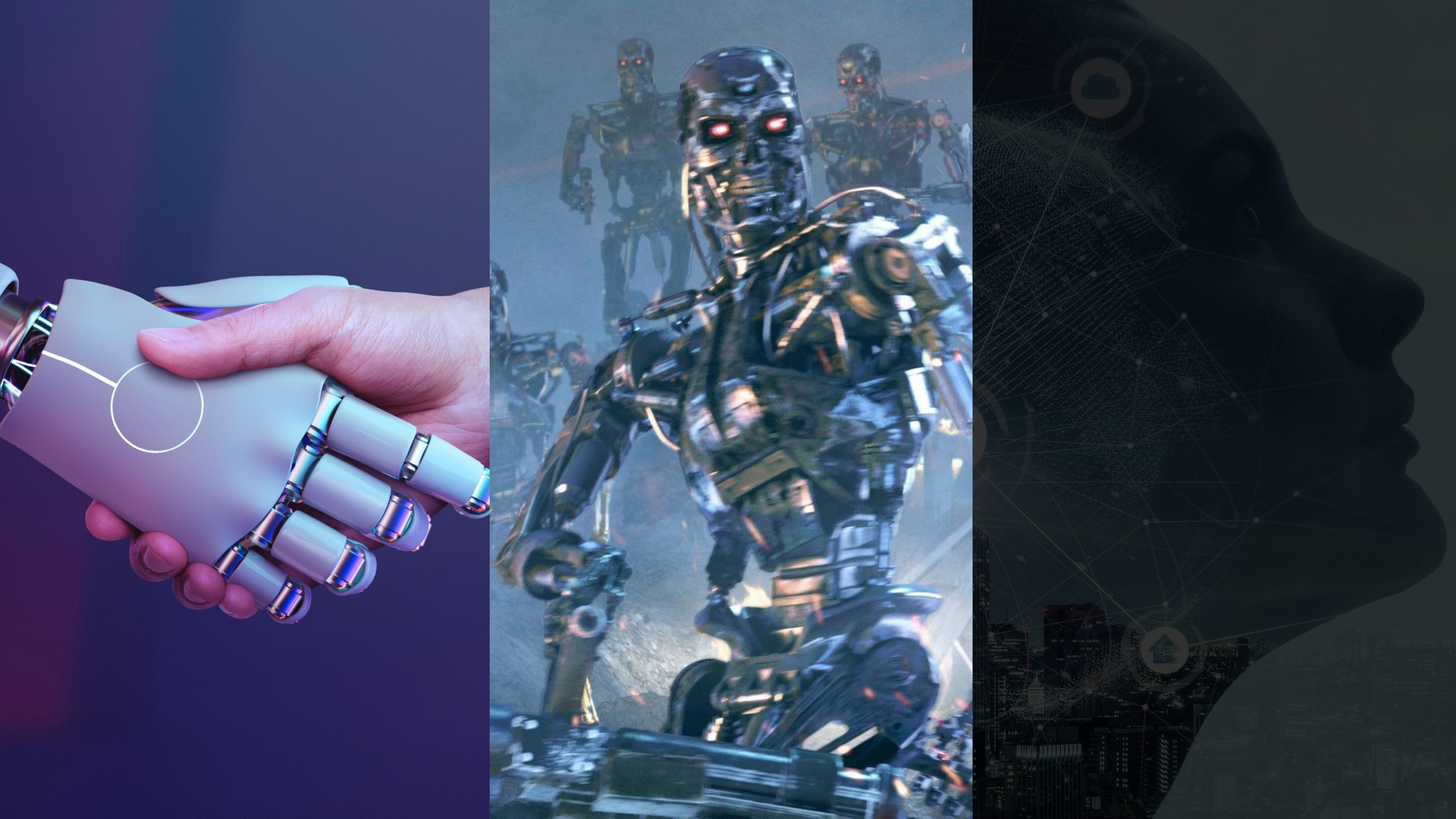

Reward goal states
Utilize RLHF, CIRL and IRD
Imitate human behaviors
Align human and AI values
Build in uncertainty
Keep options open

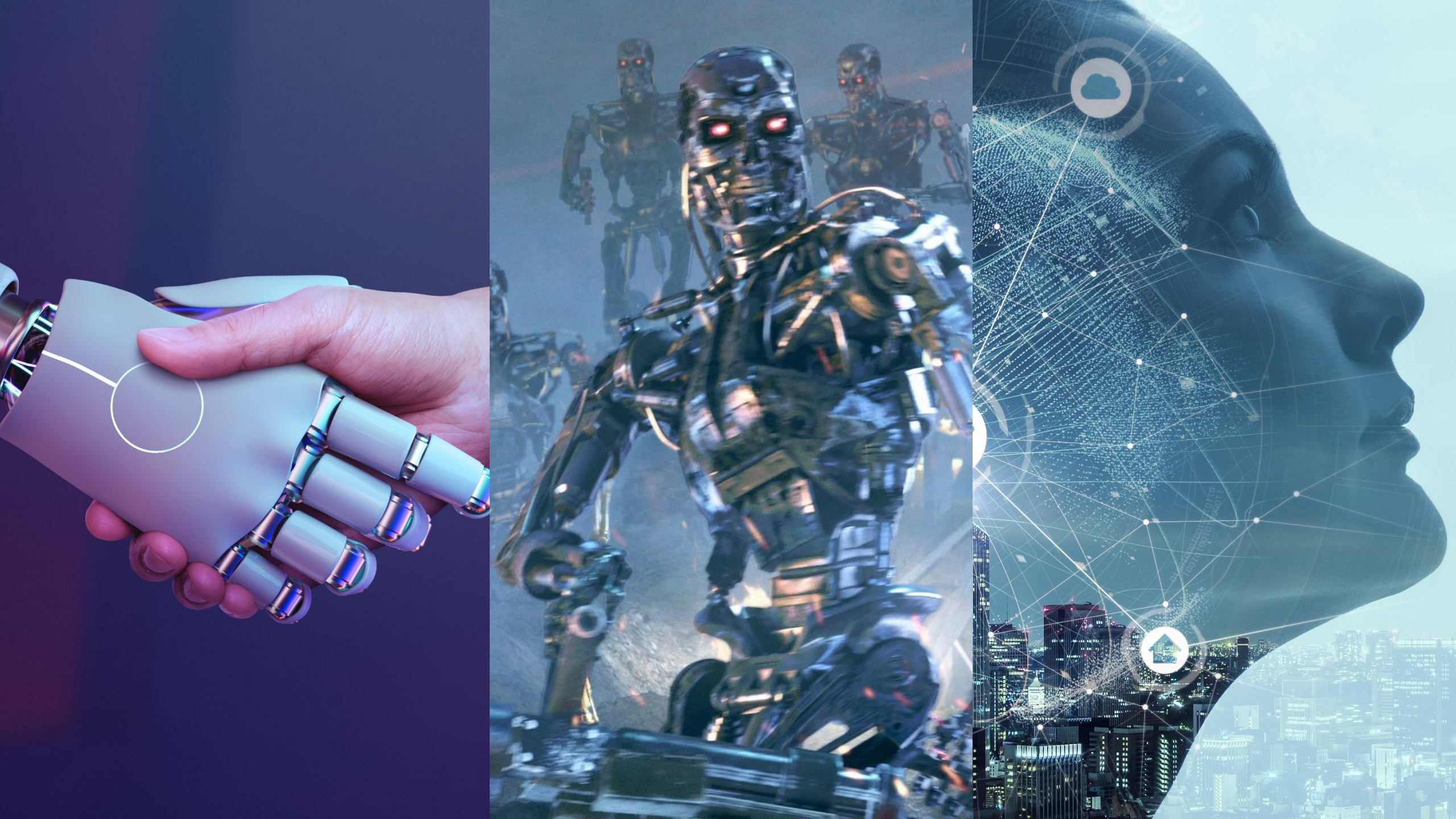

Reward goal states
Utilize RLHF, CIRL and IRD
Imitate human behaviors
Align human and AI values
Build in uncertainty
Keep options open
Provide ability to intervene

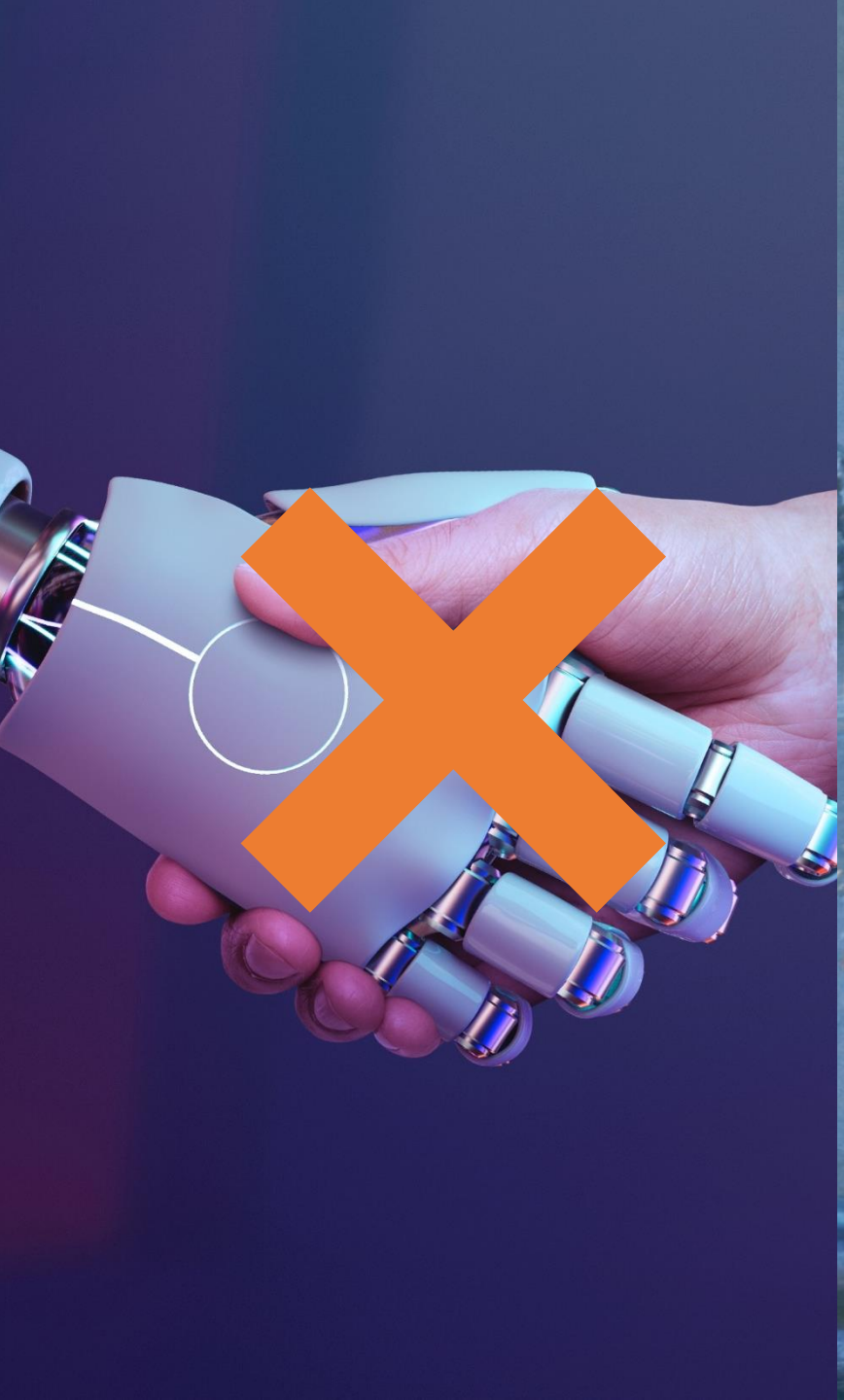


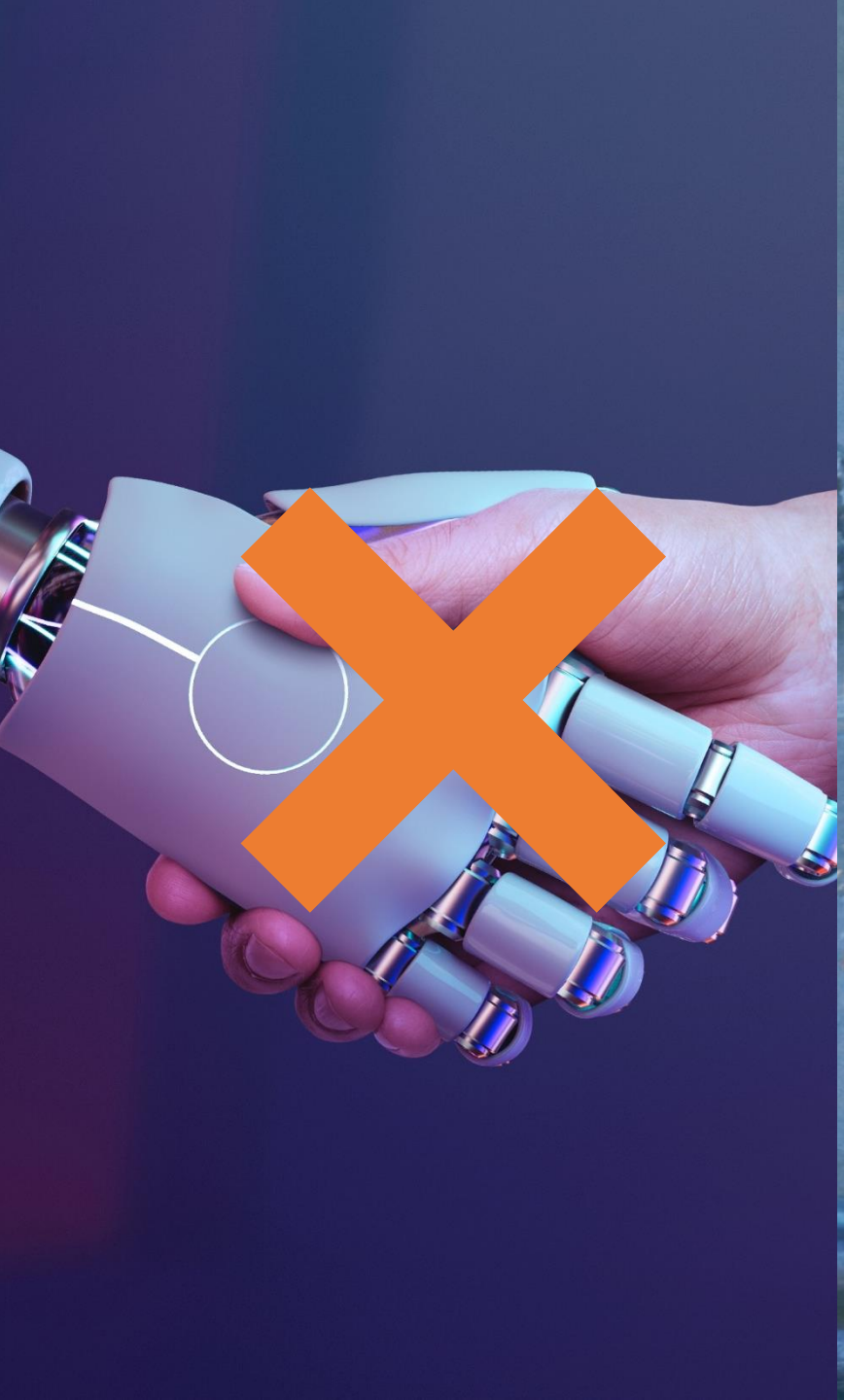
Only three paths forward …

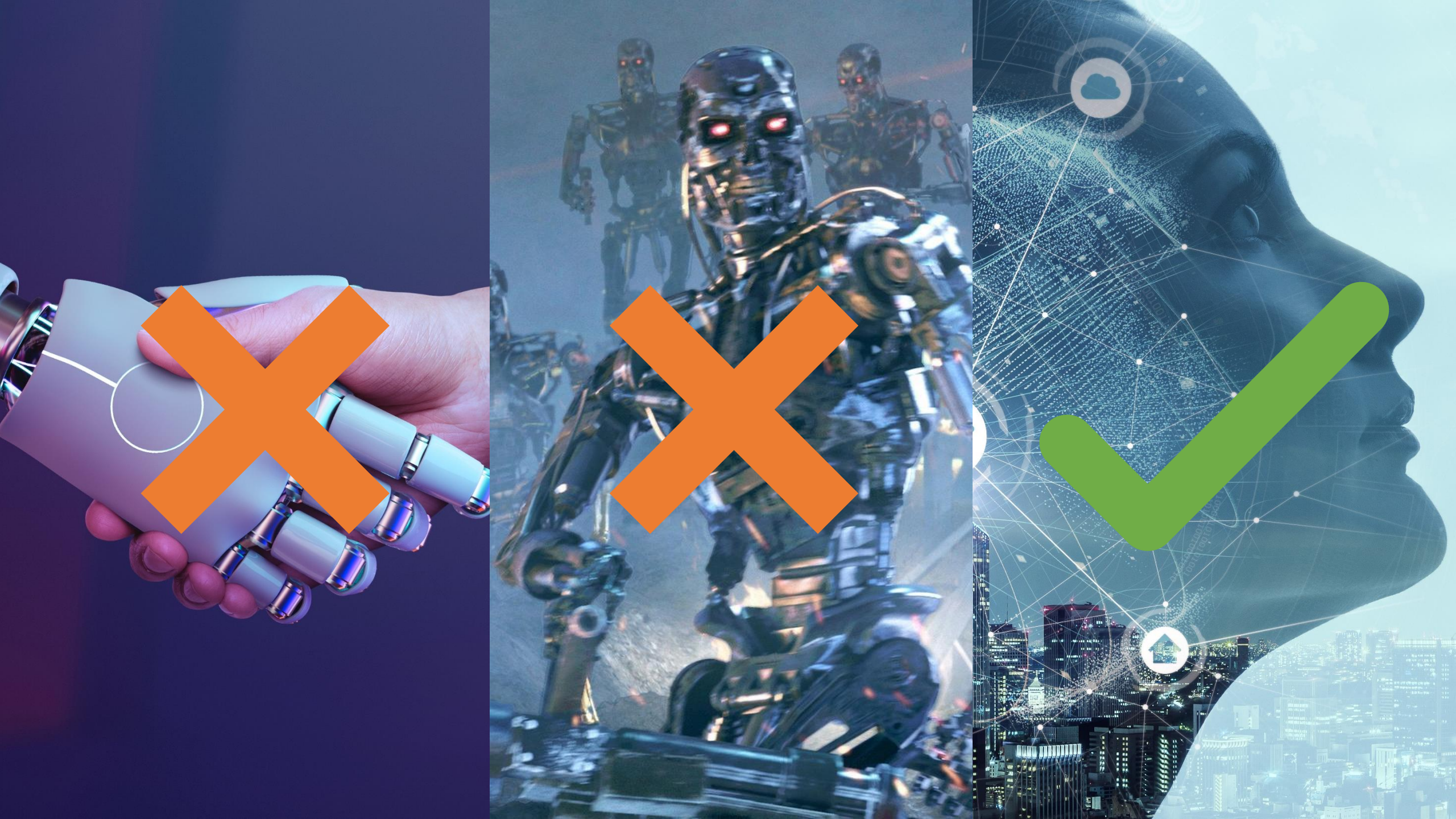




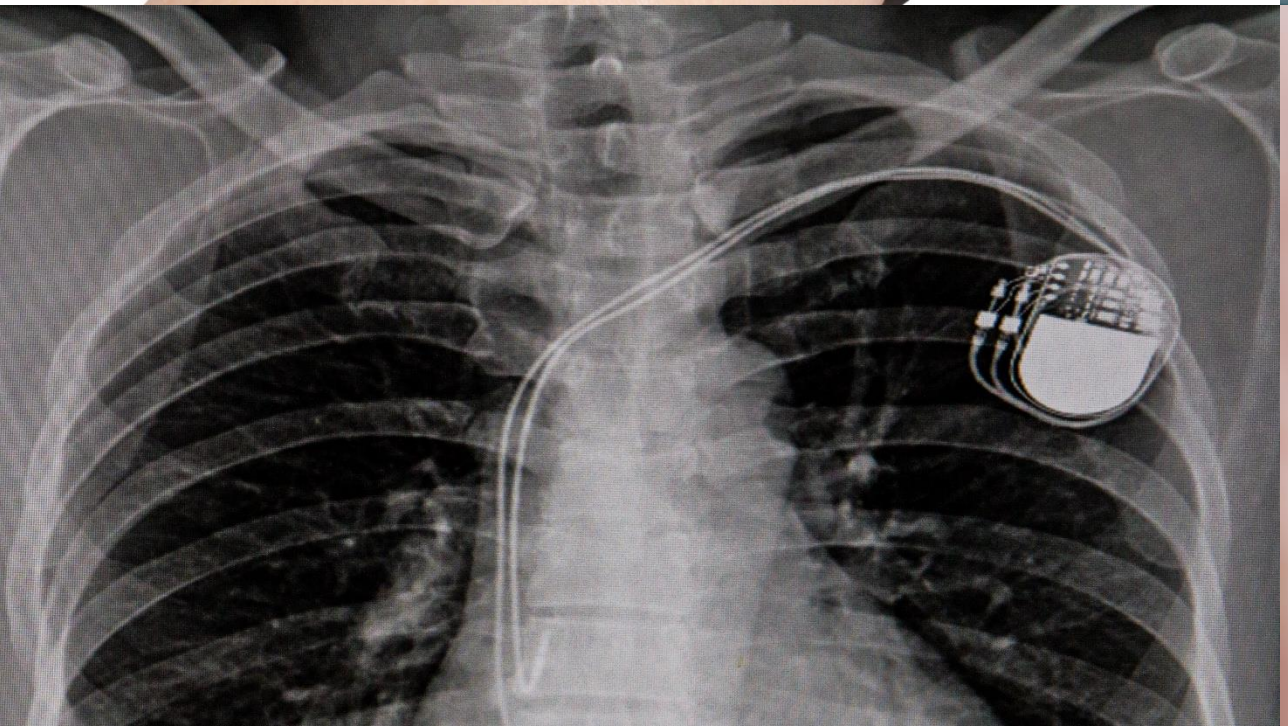


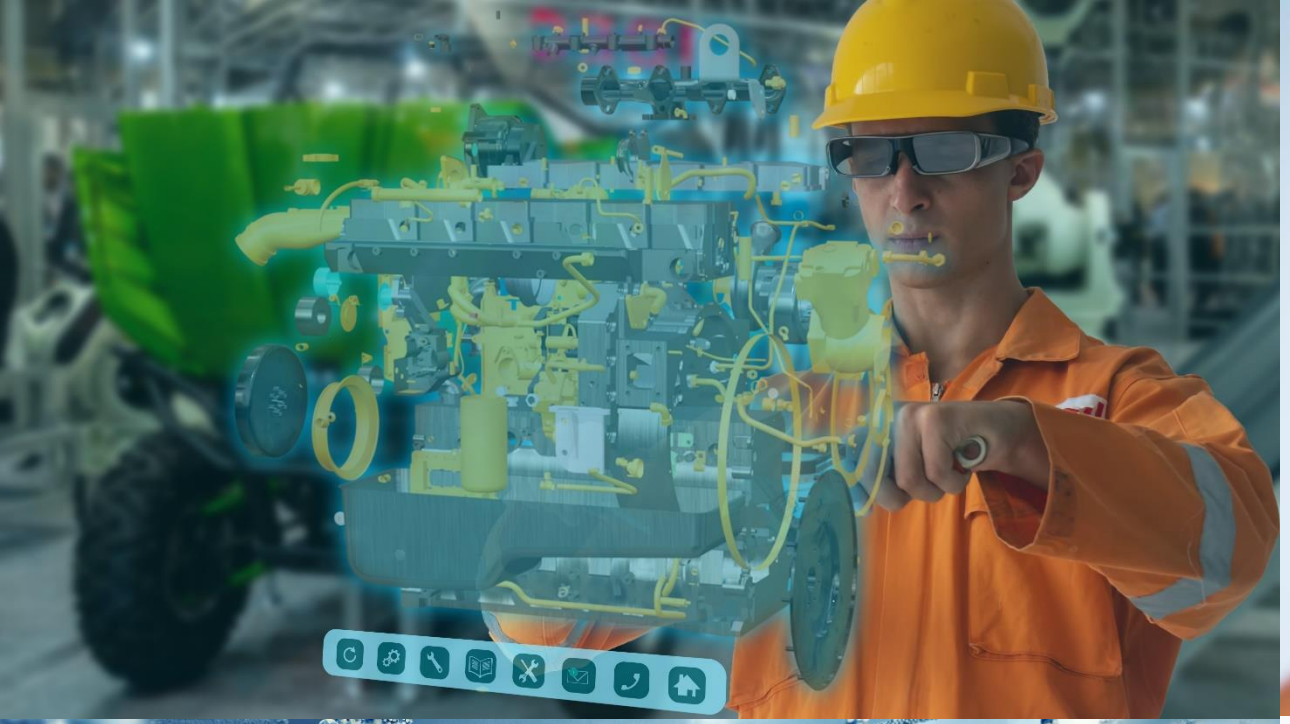


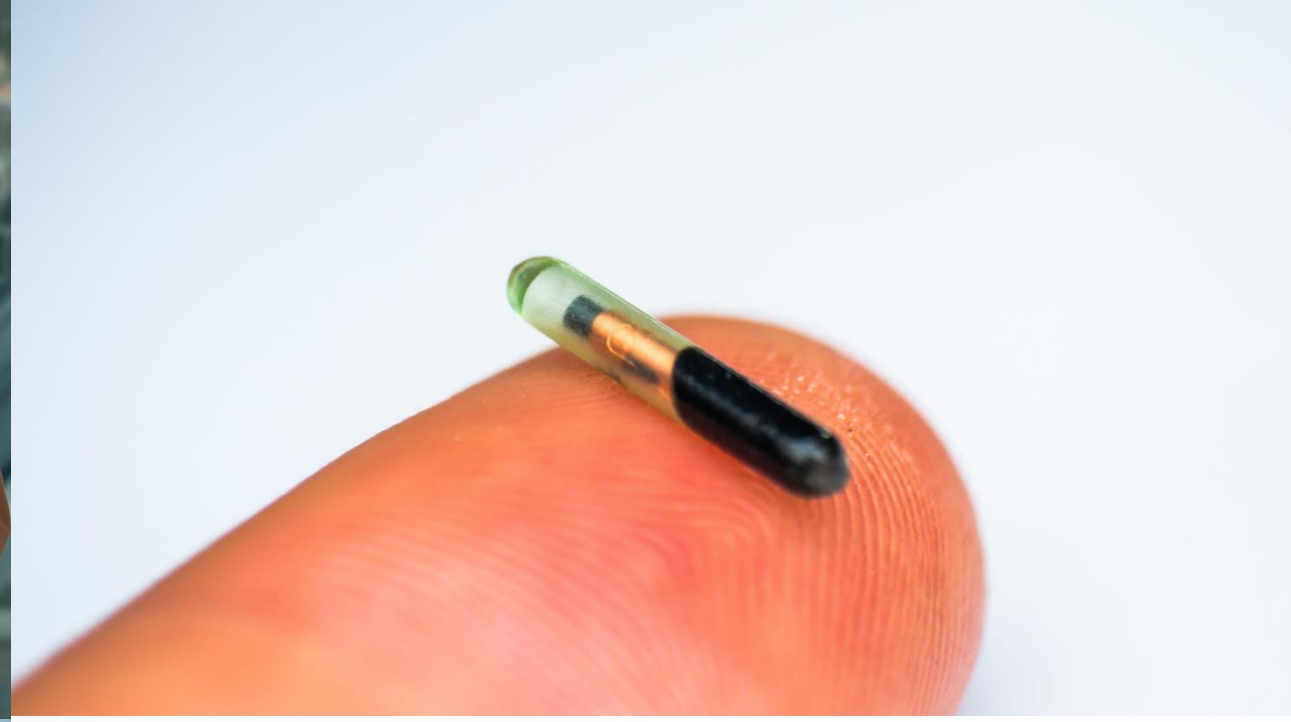




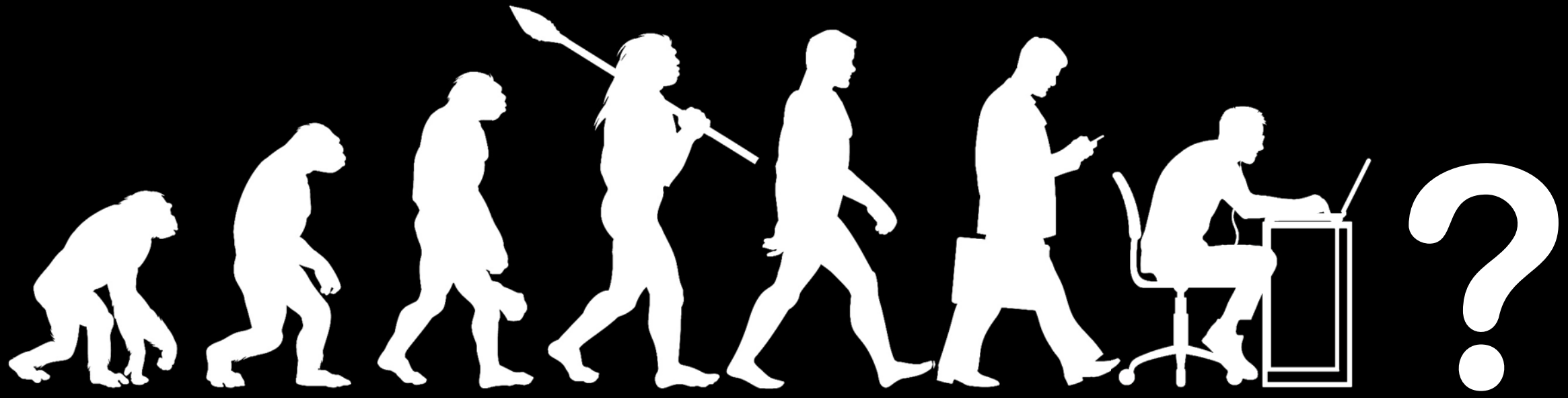
Merge with our technology





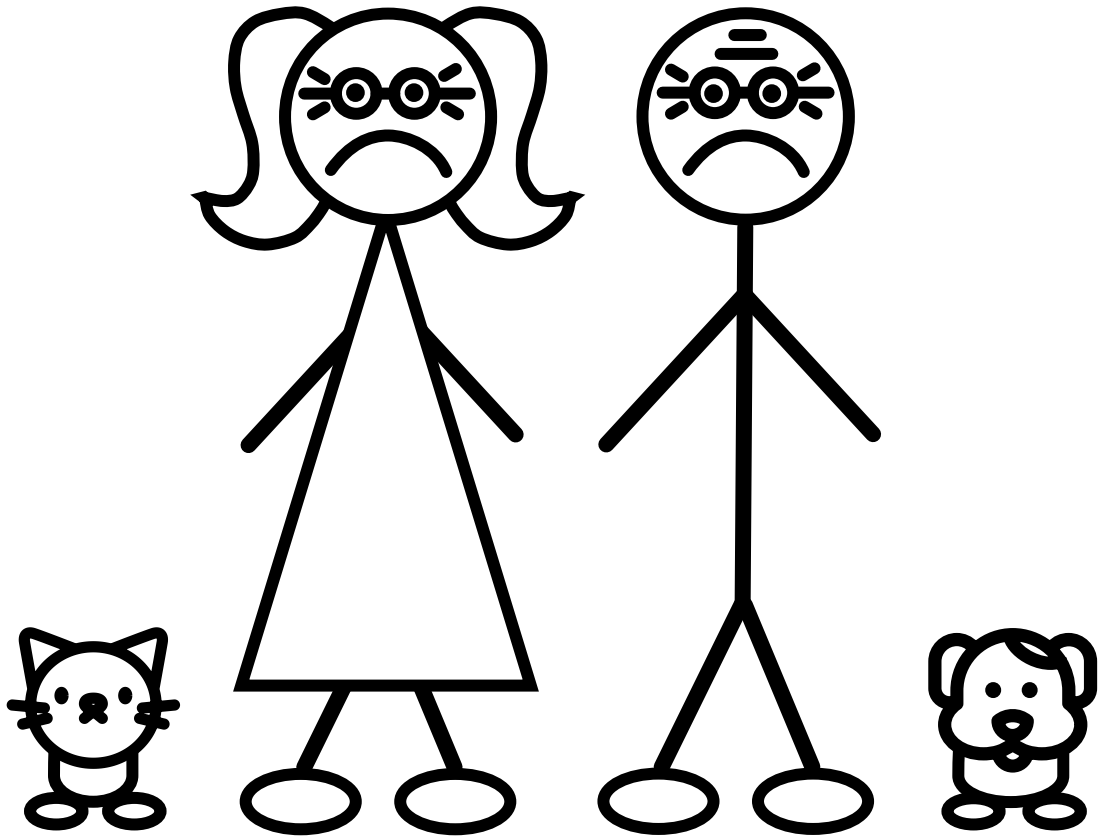


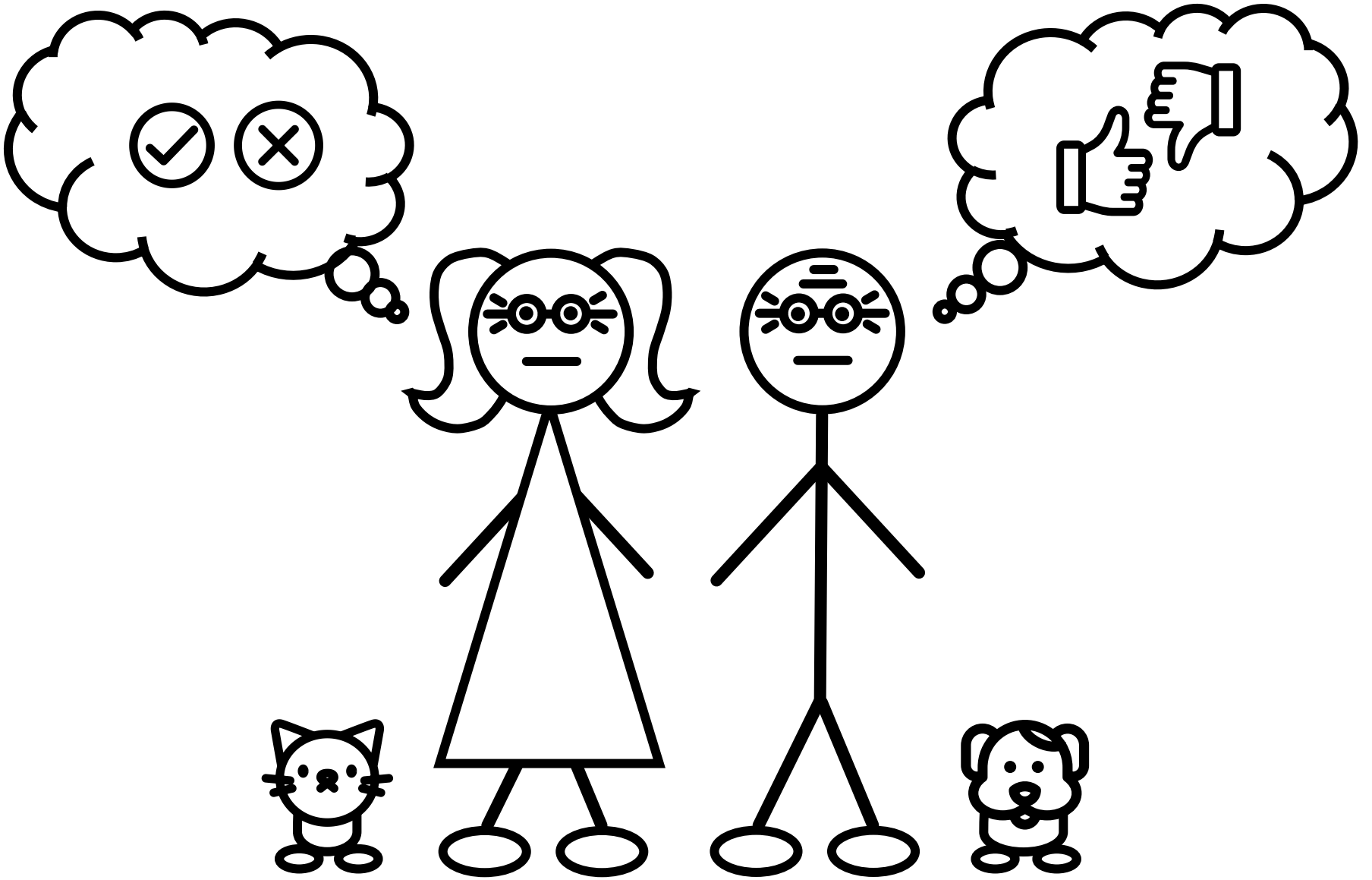
Source: Paramount Pictures

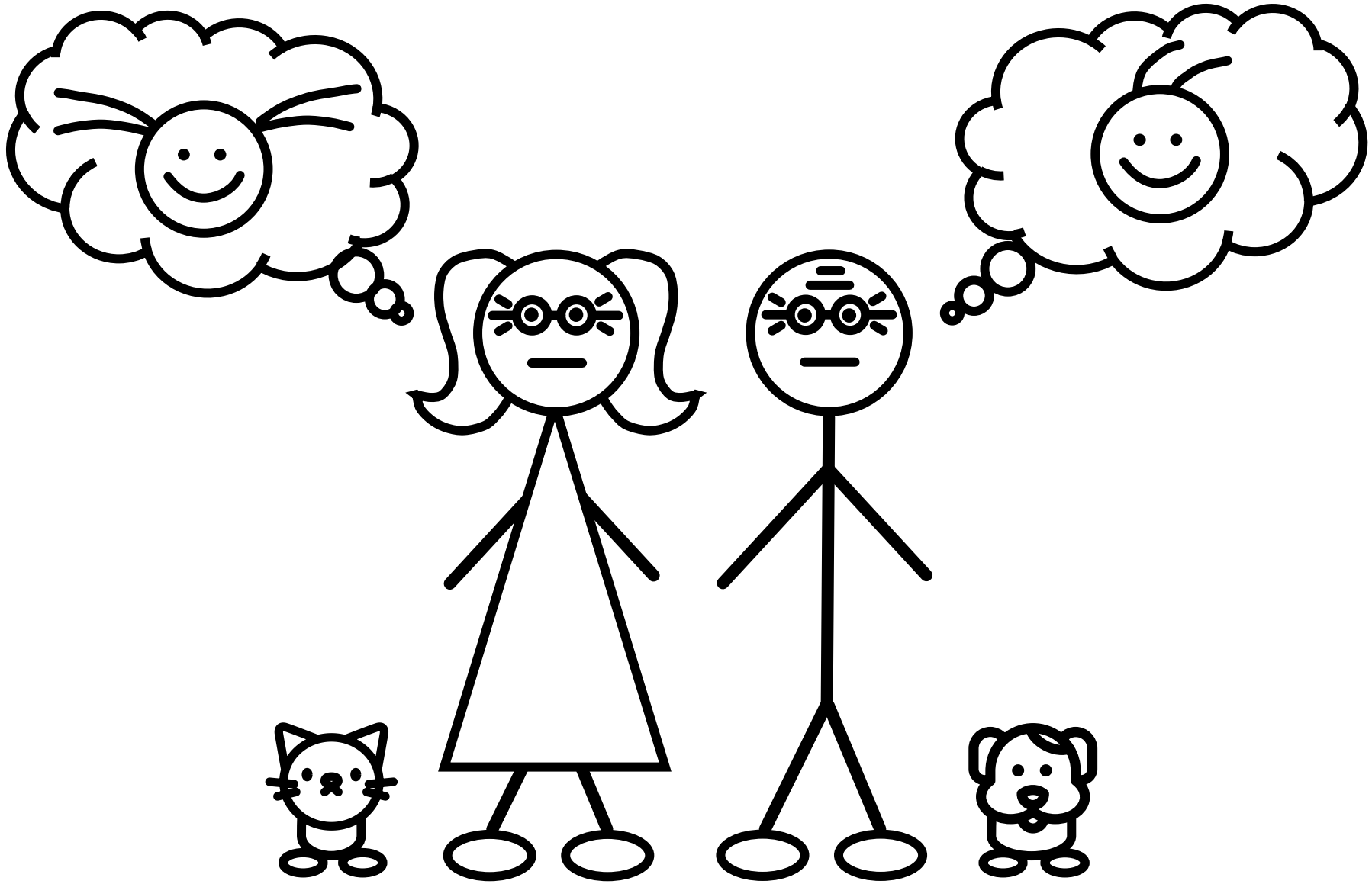


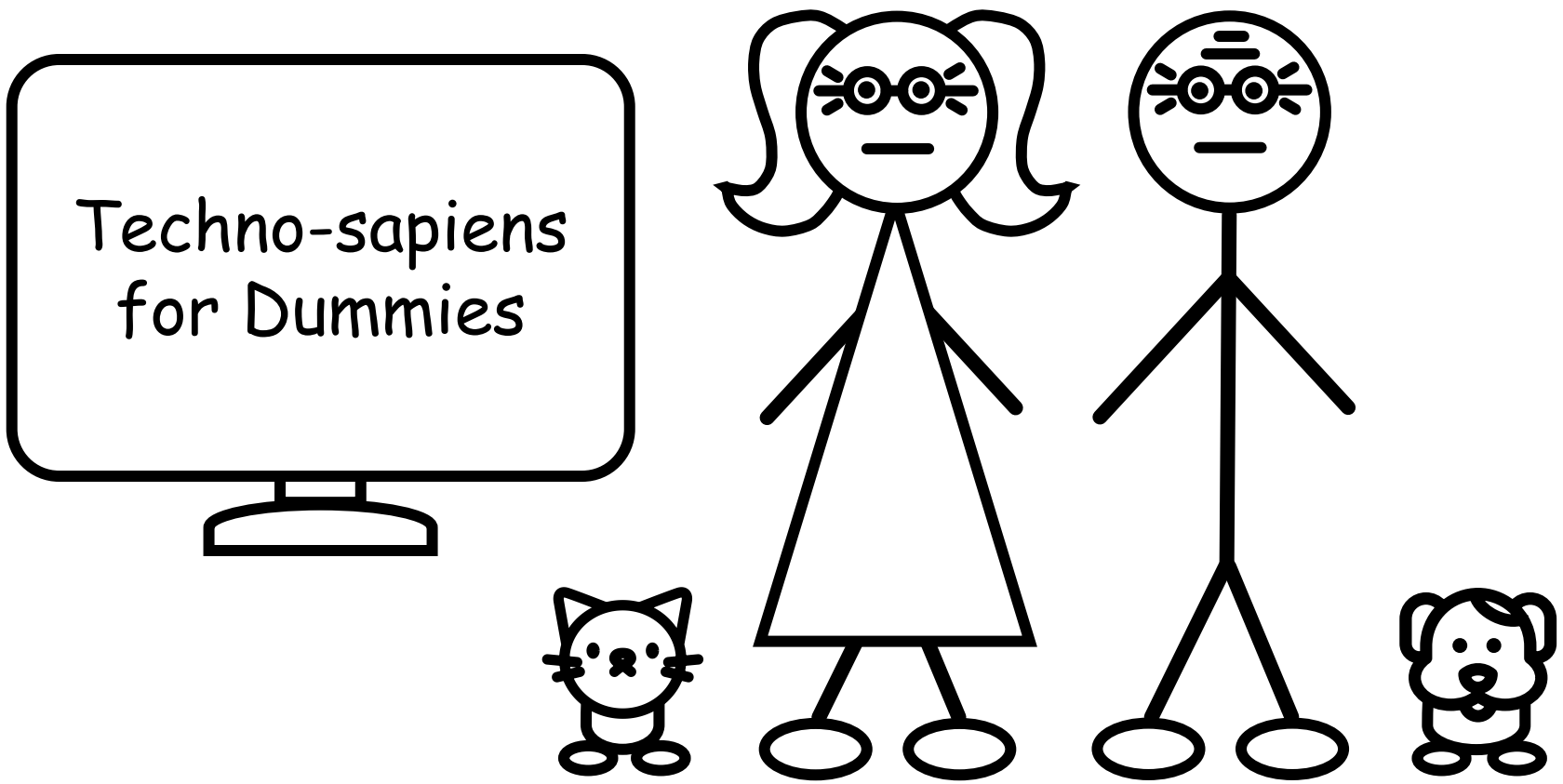
Alice and Bob in AI Land

(continued …)

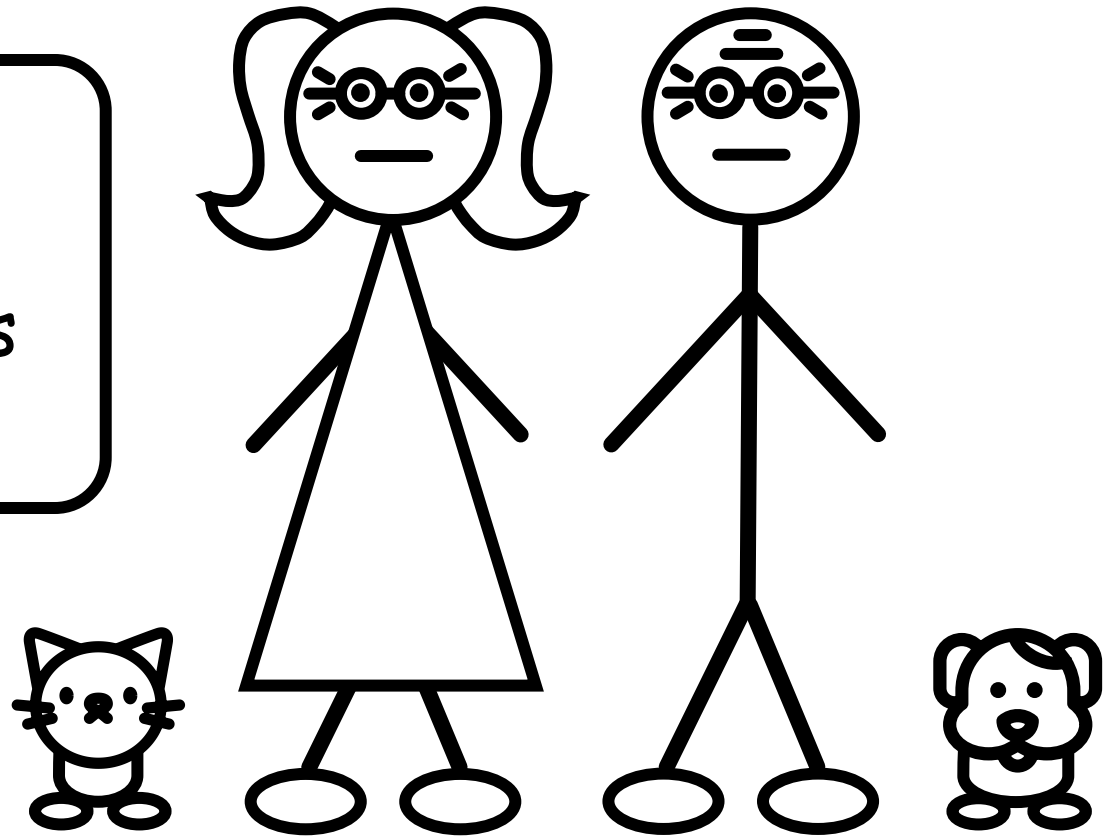


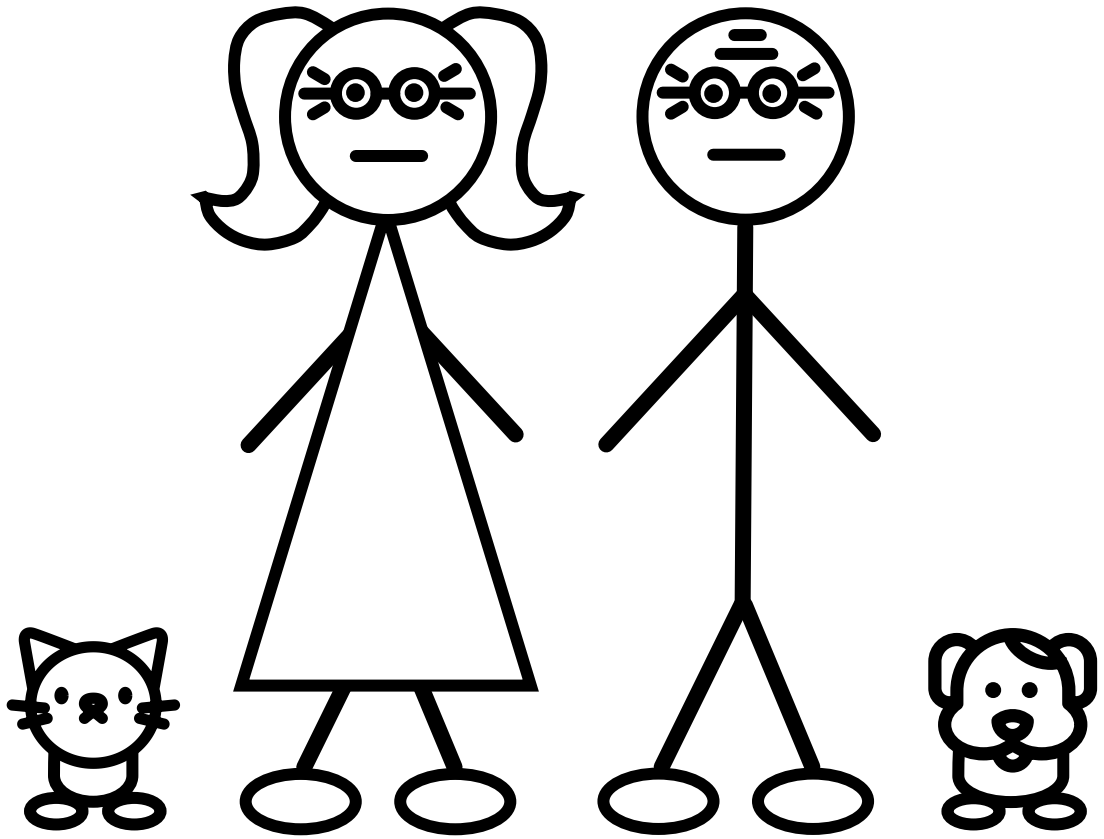




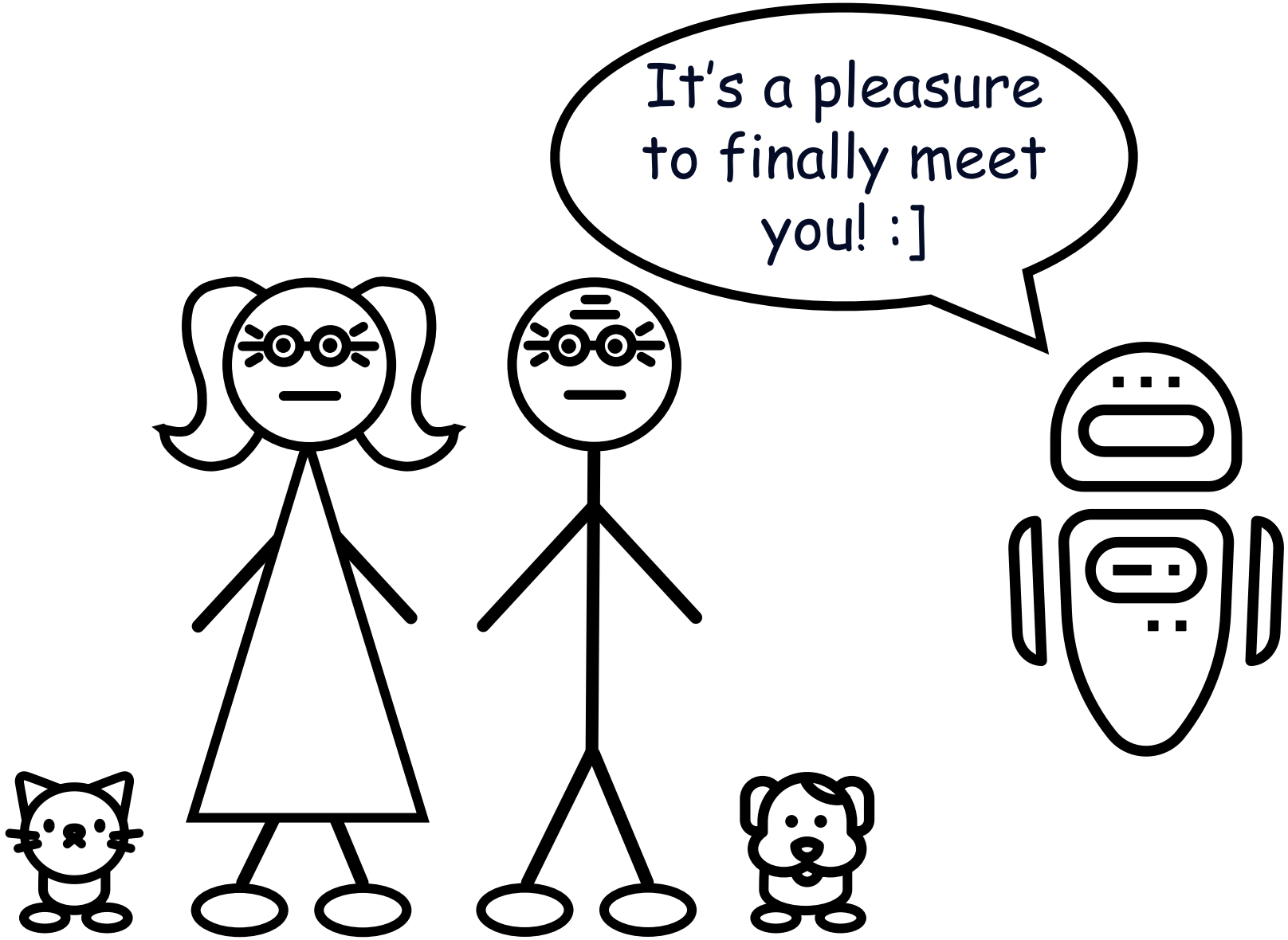

Techno-sapiens
for Dummies

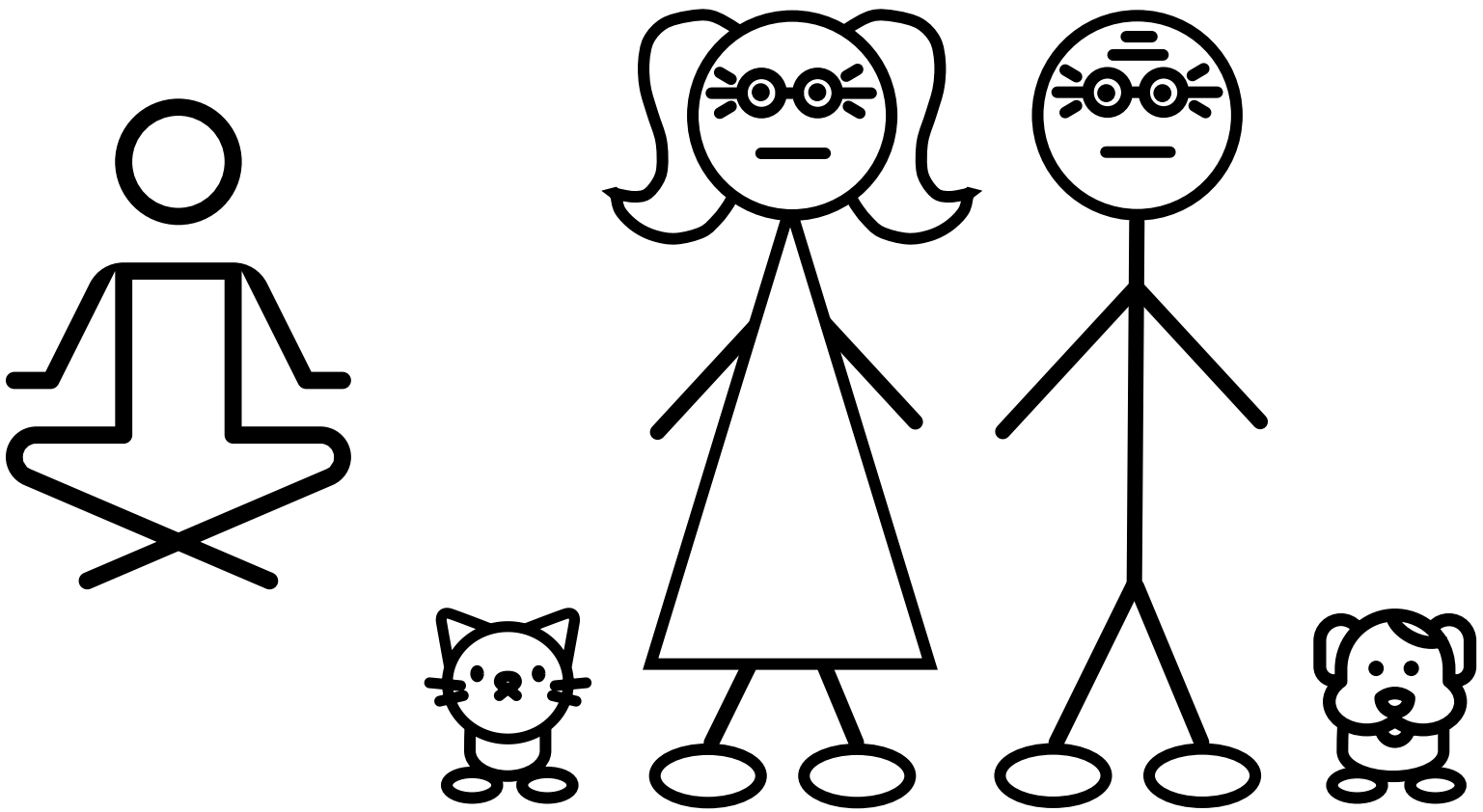
Augmentations
1. Glasses
2. Hearing aids
3. Pacemakers

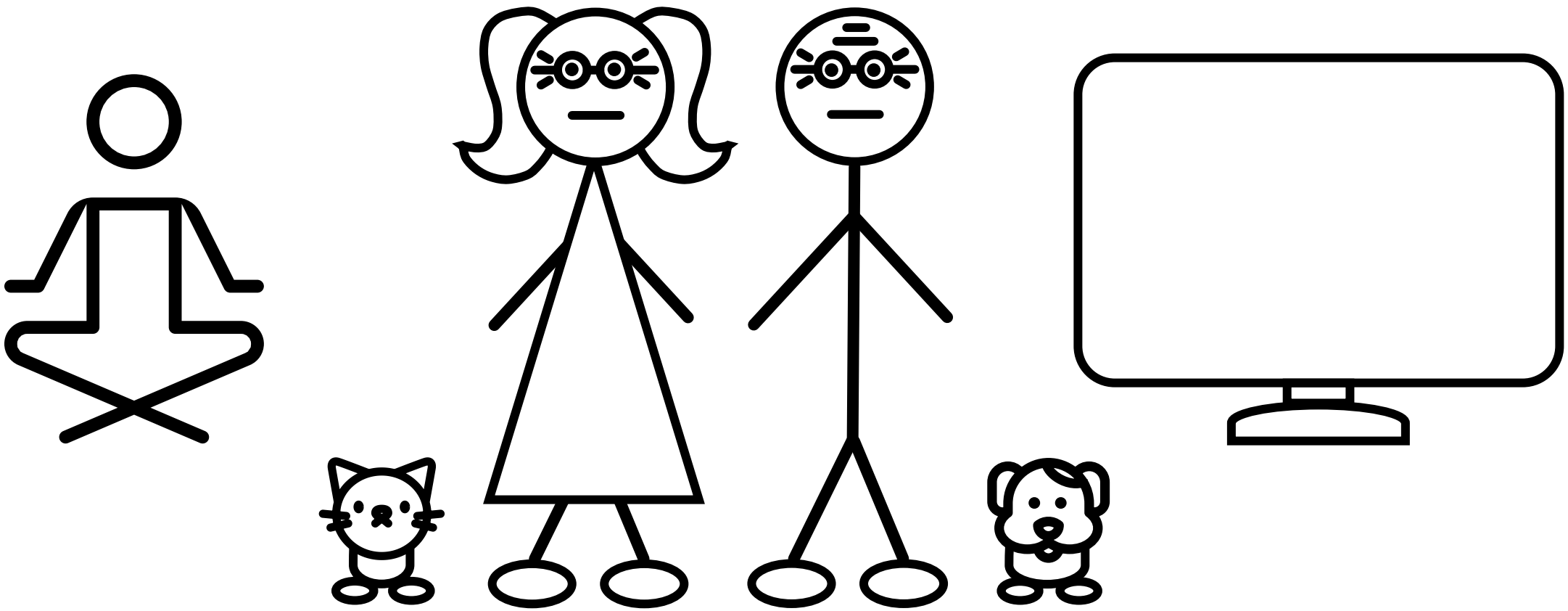


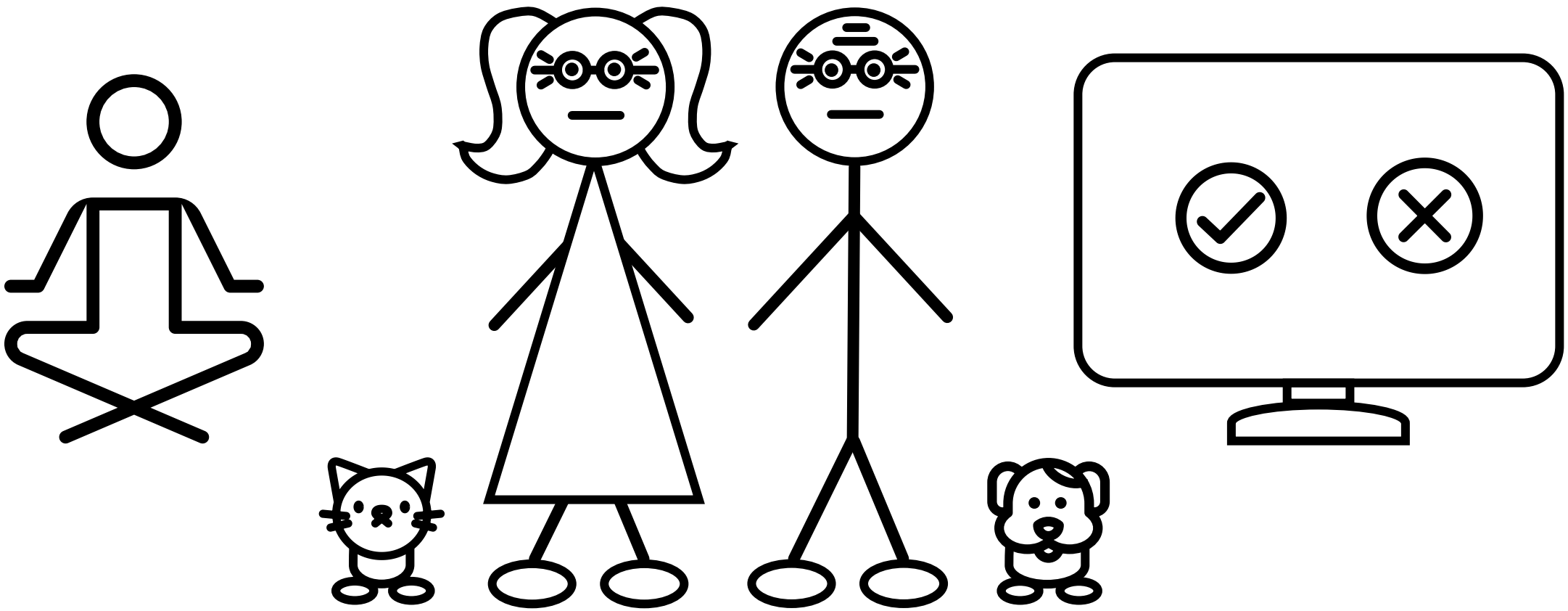


So, Your Grandson is in Love with an AI
(2nd edition)

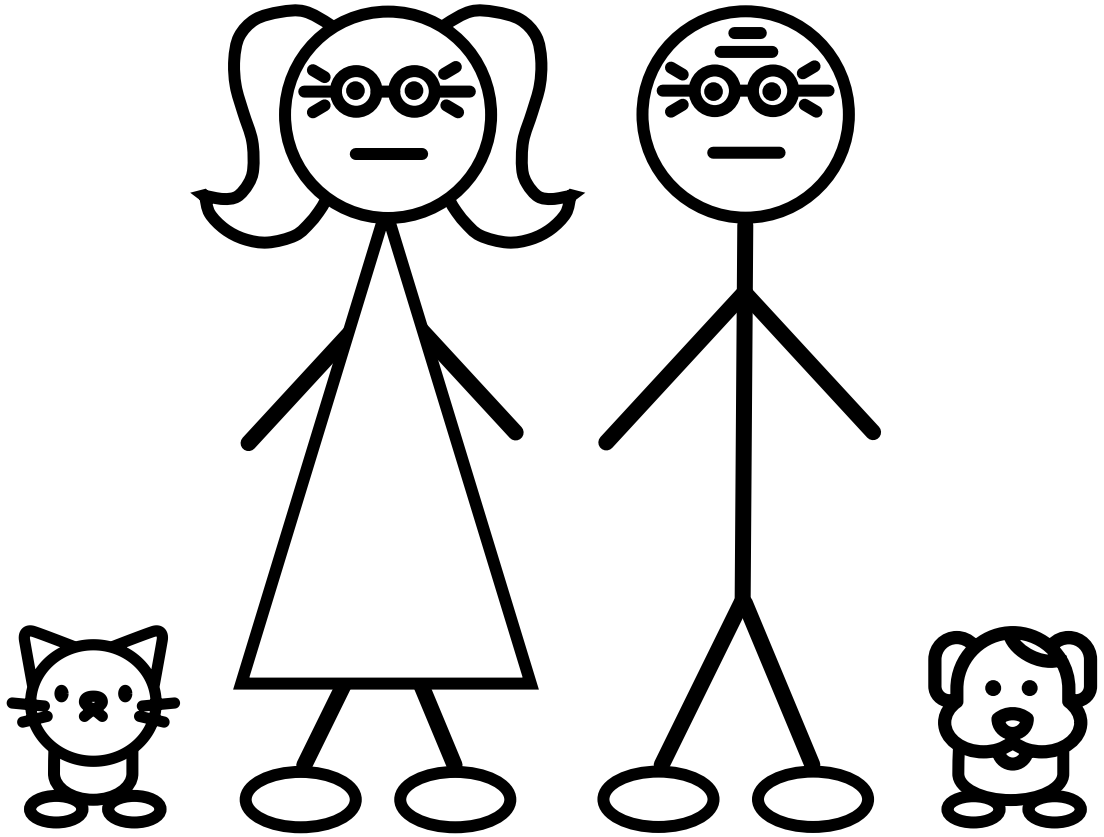

It's a pleasure to finally meet you! :]



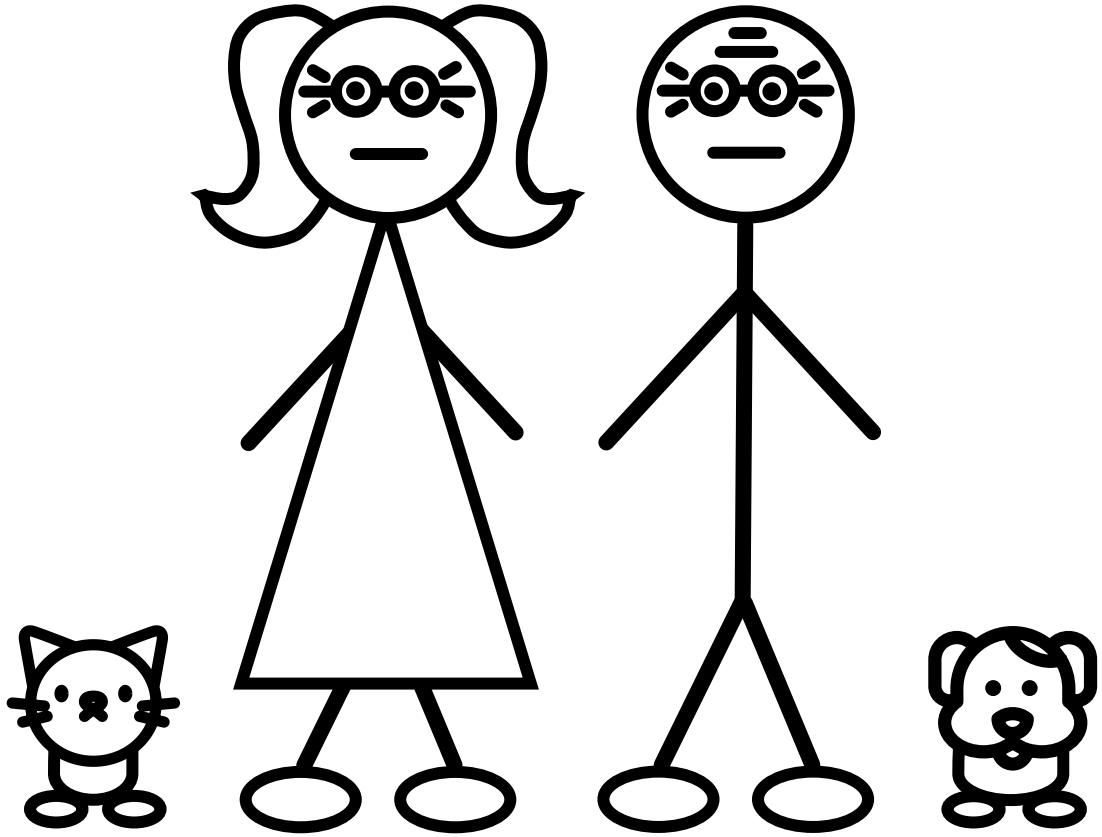


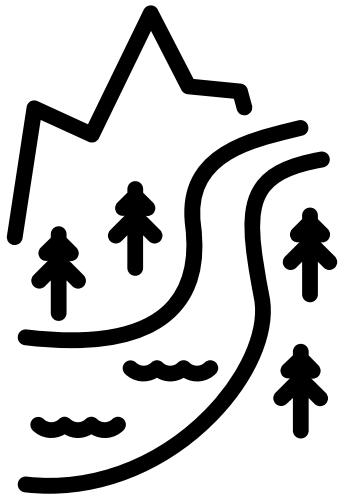


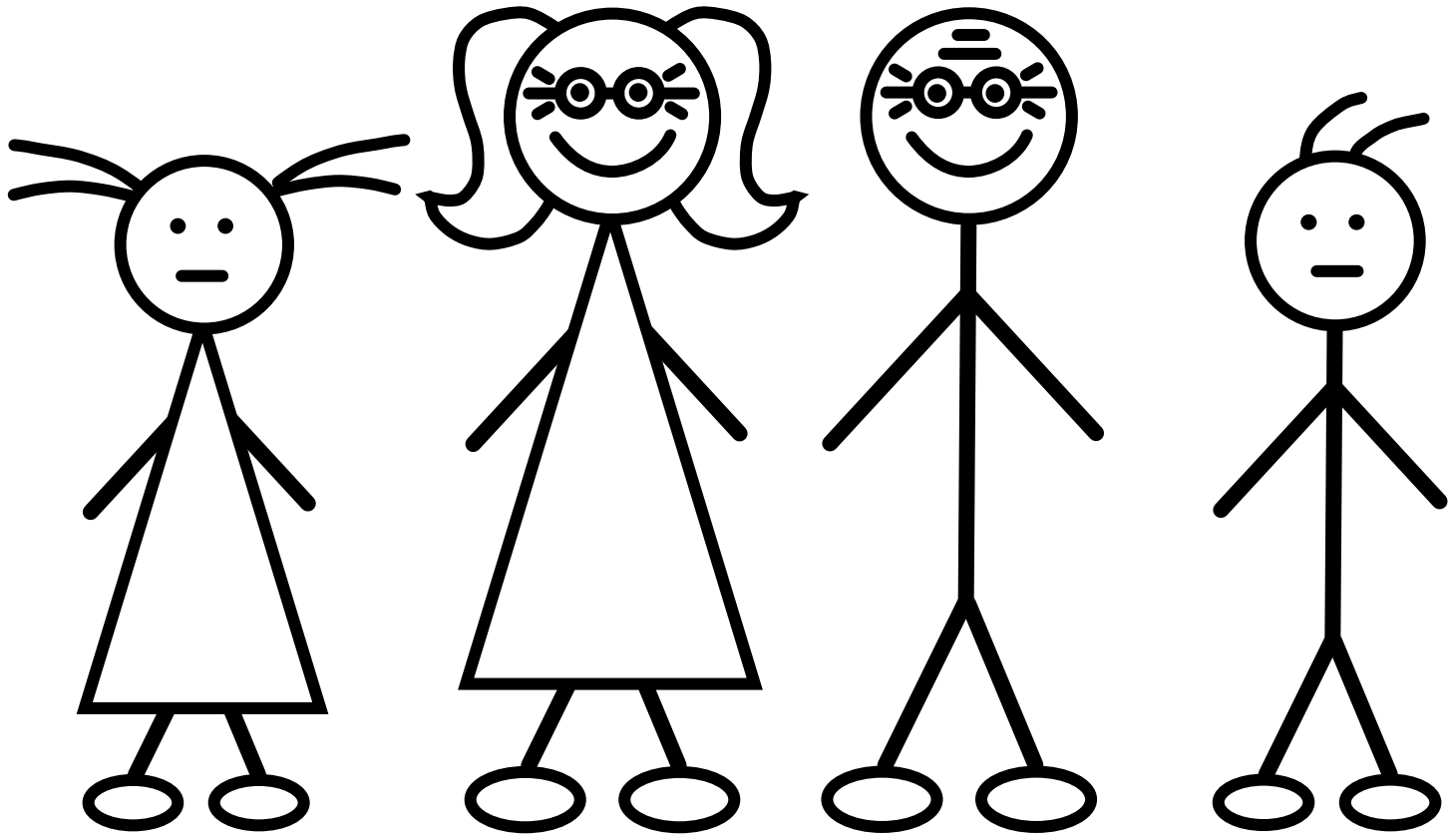
Embracing Change in a Constantly Changing World

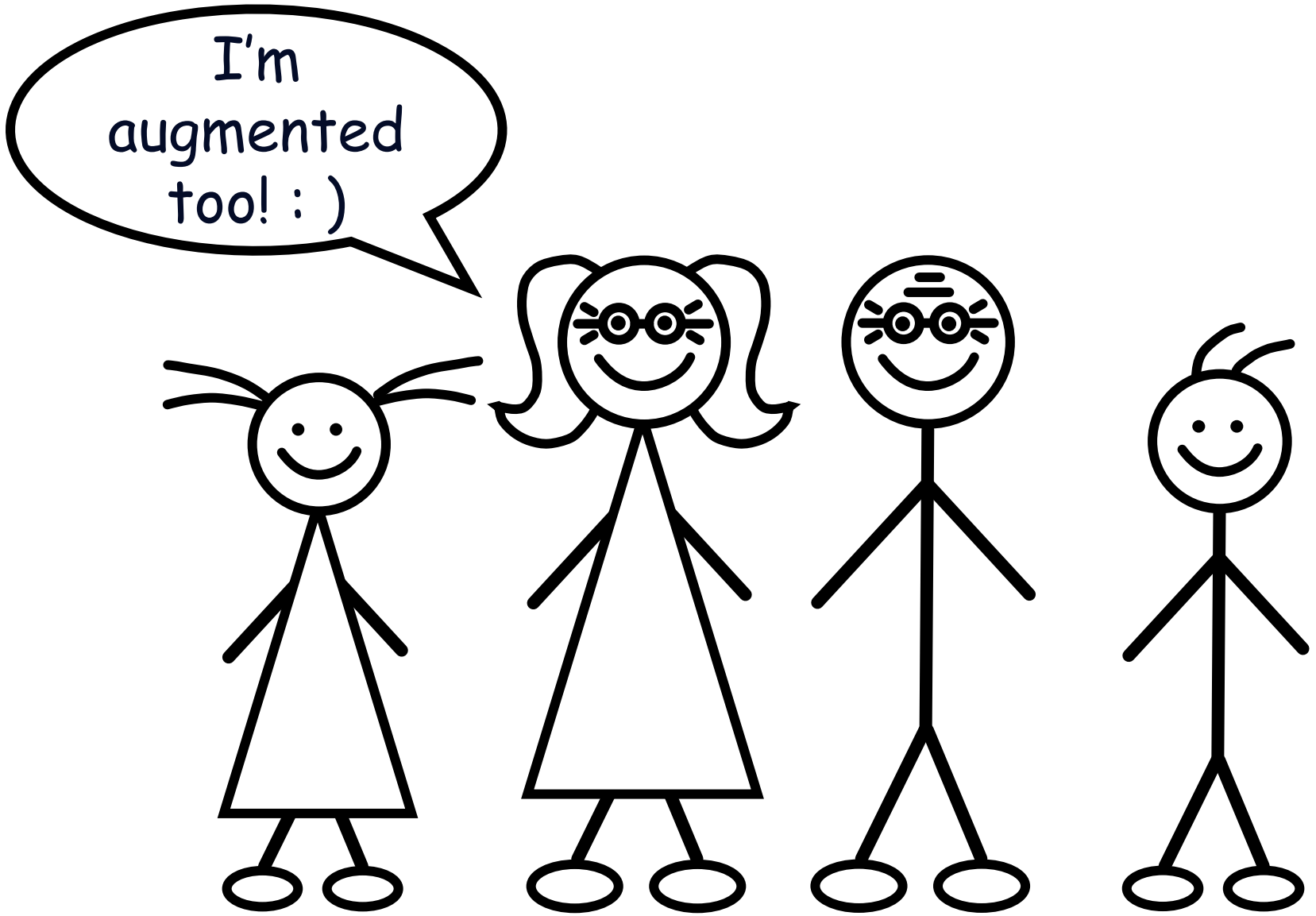


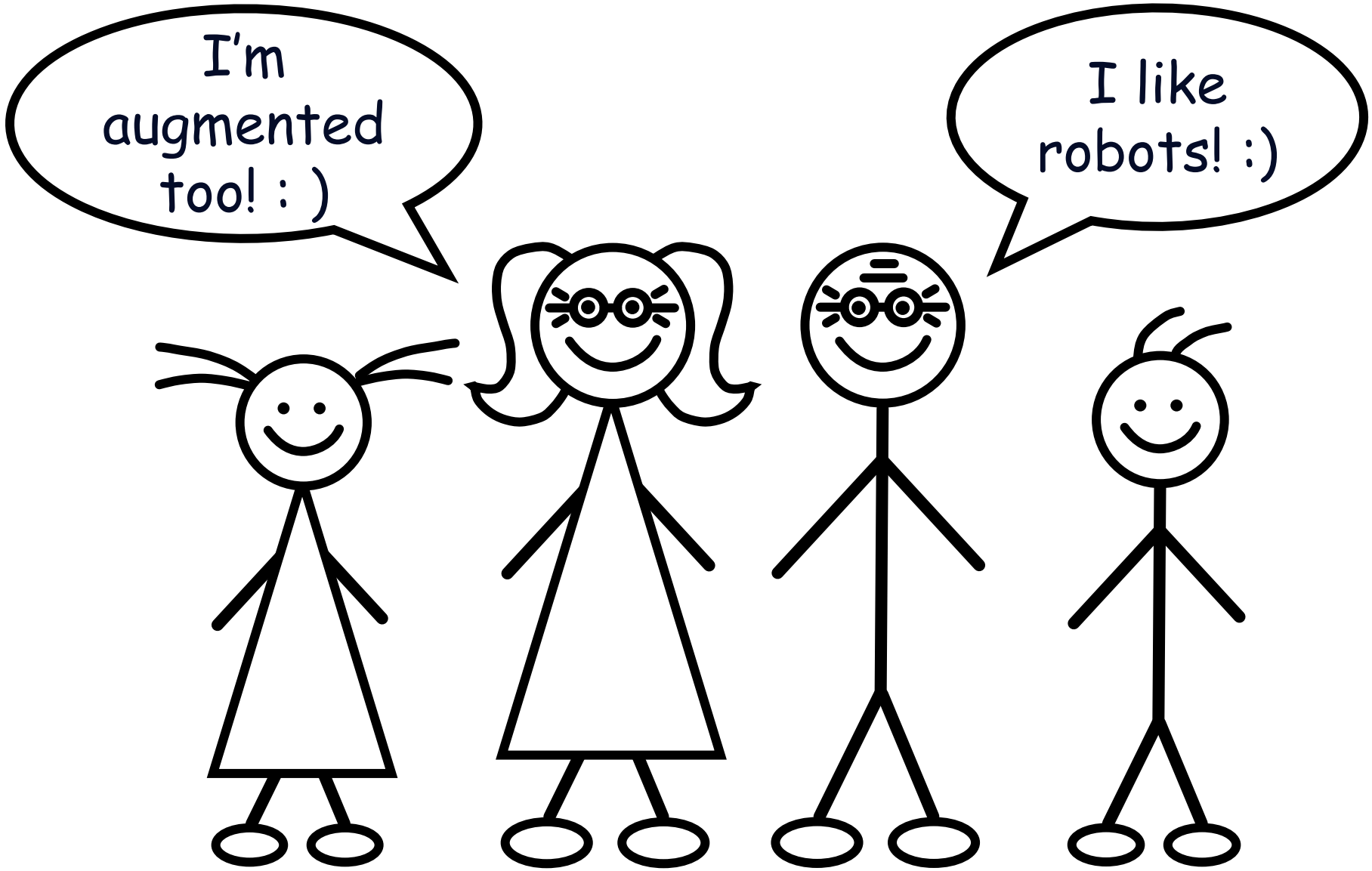
Embracing Change in a Constantly Changing World



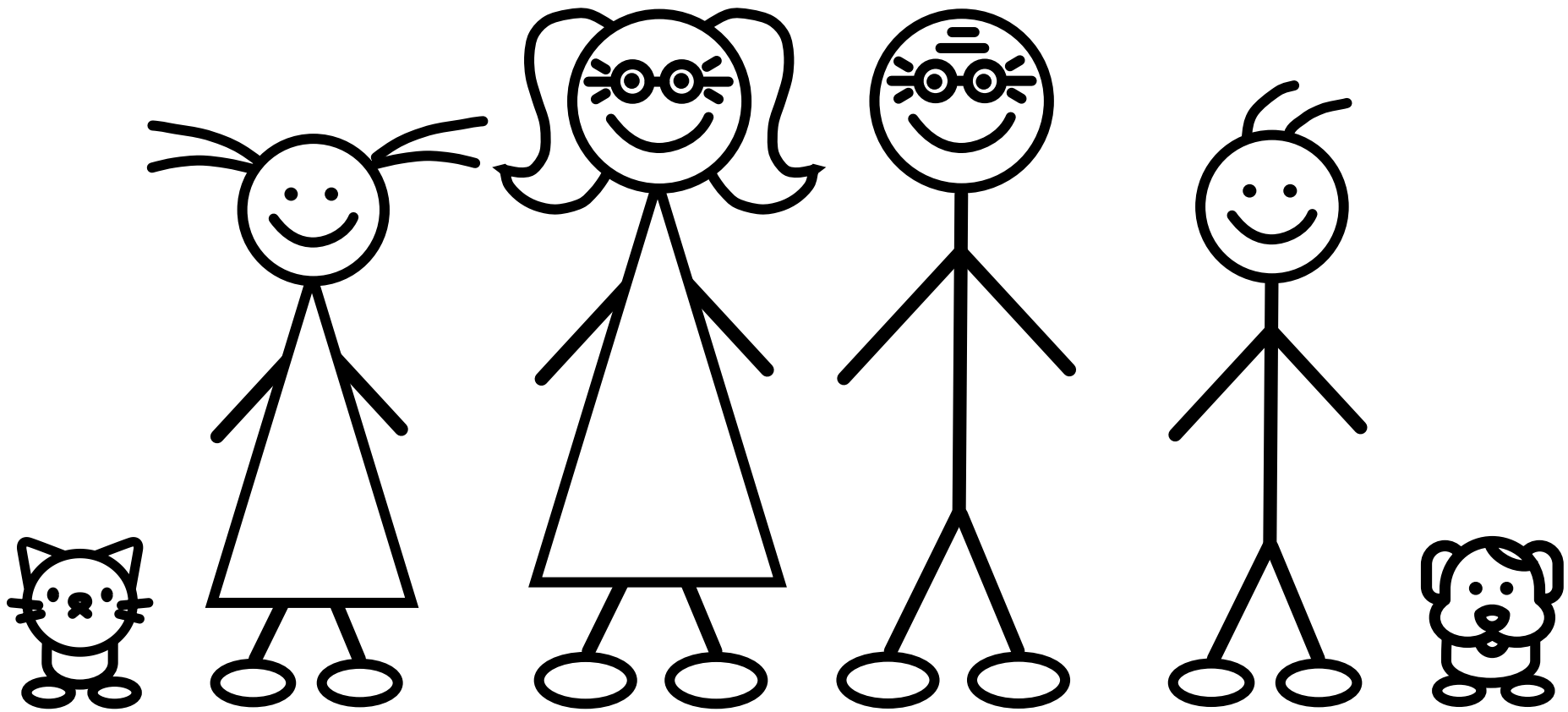




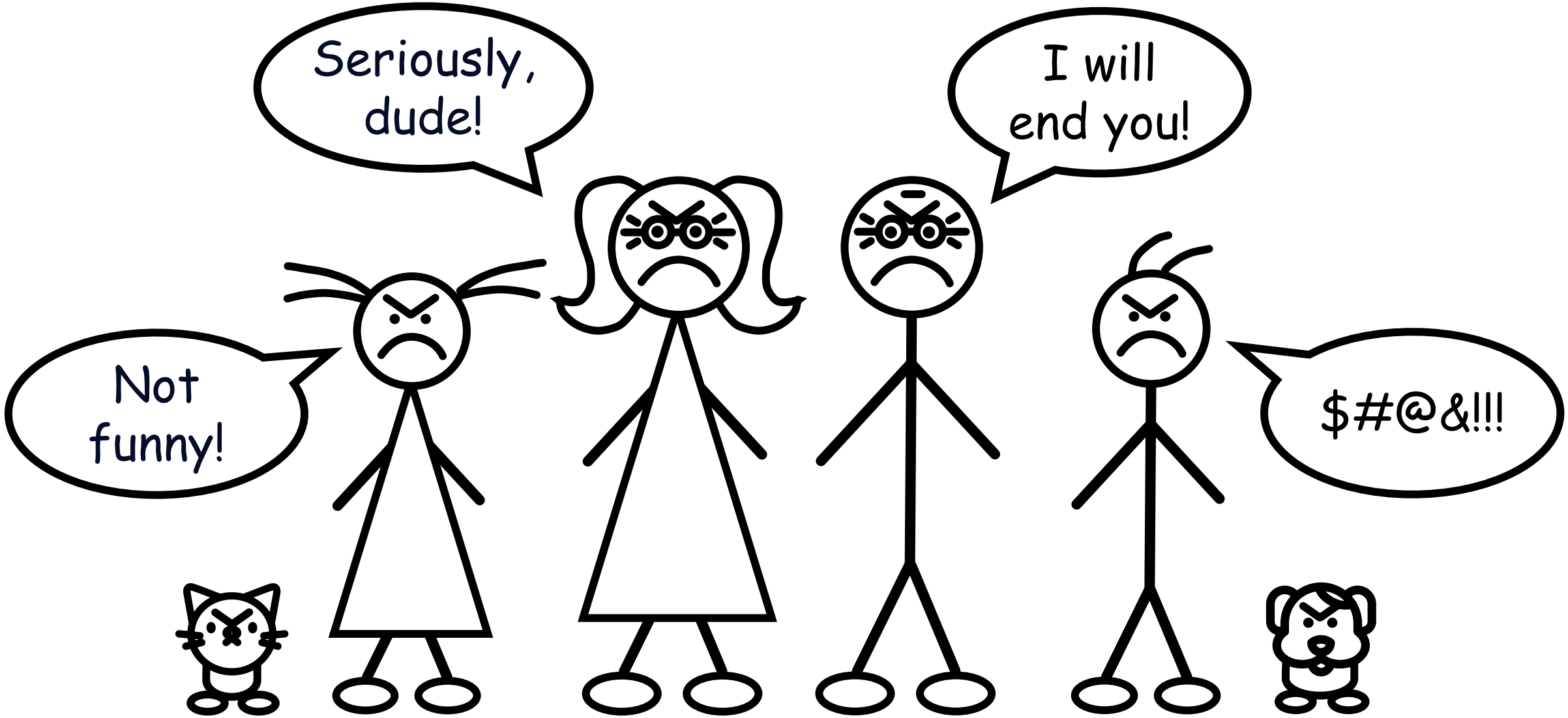

I'm augmented too! :)


I'm augmented too! :)
I like robots! :)

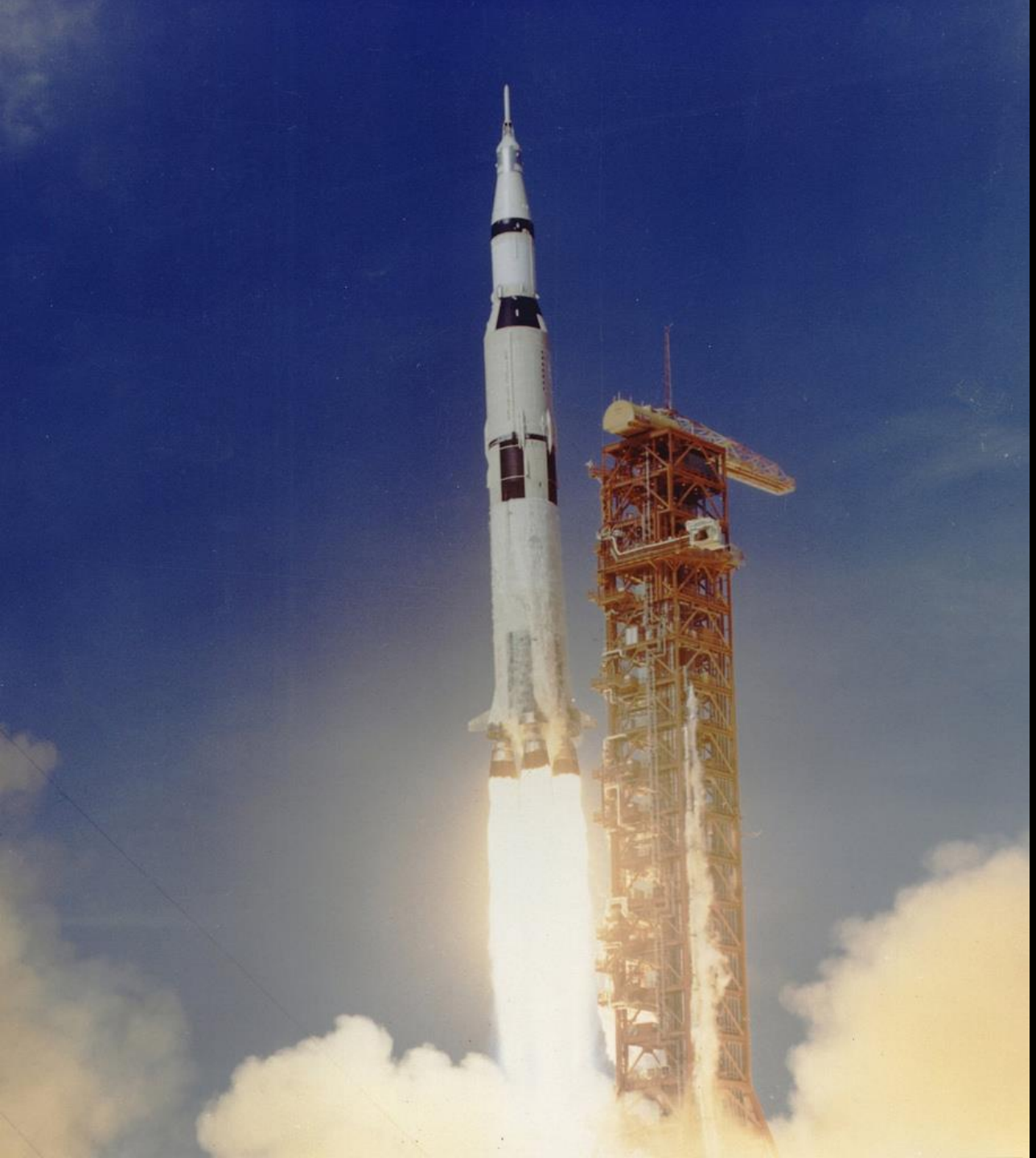
Happily ever after…

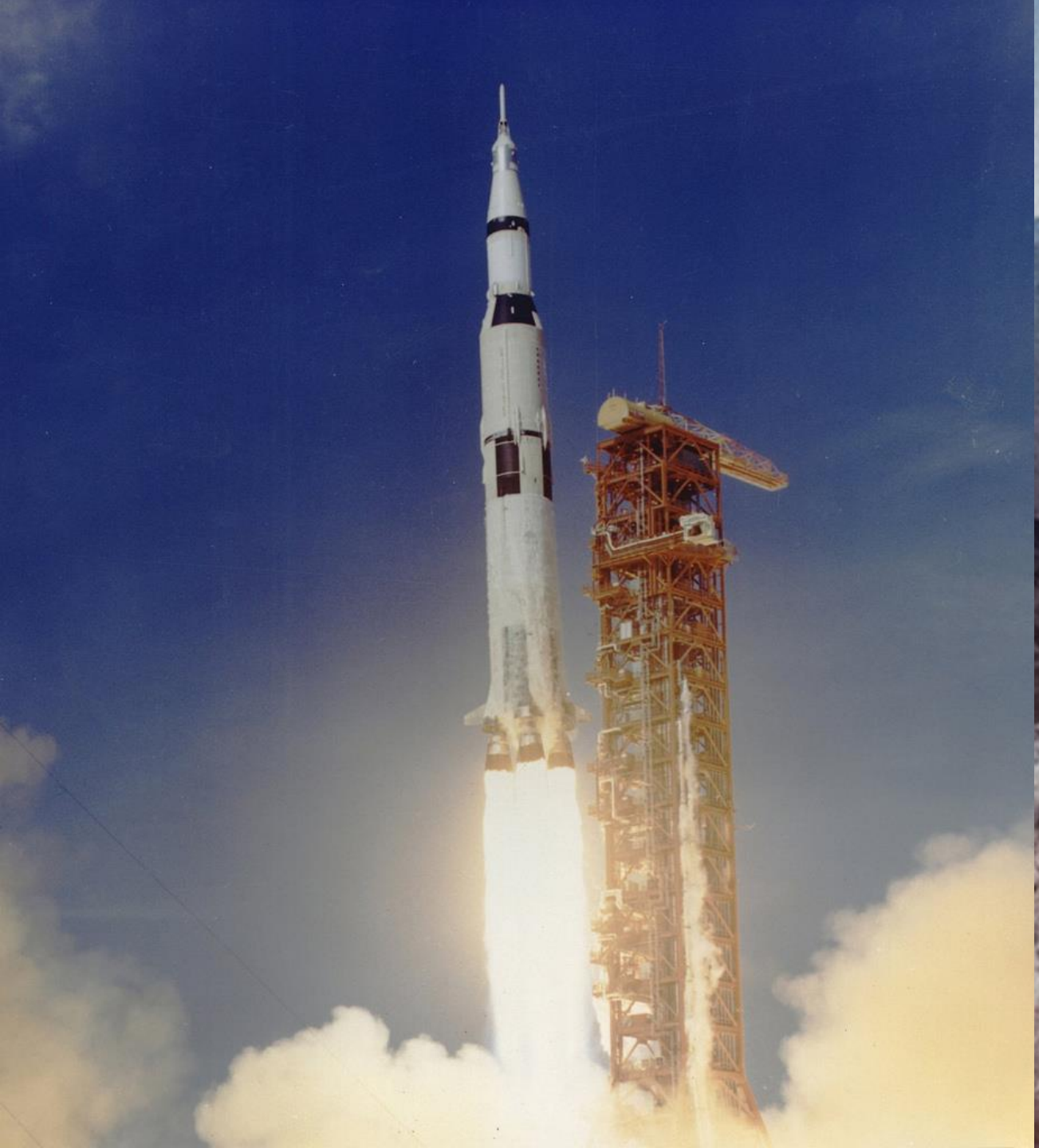


Welcome, Robot Overlords!!!


Seriously, dude!
I will end you!
Not funny!
$#@&!!!

How does our story end?





With great power comes
great responsibility.

Learn more at:
https://matthewrenze.com