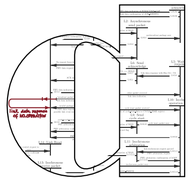
Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

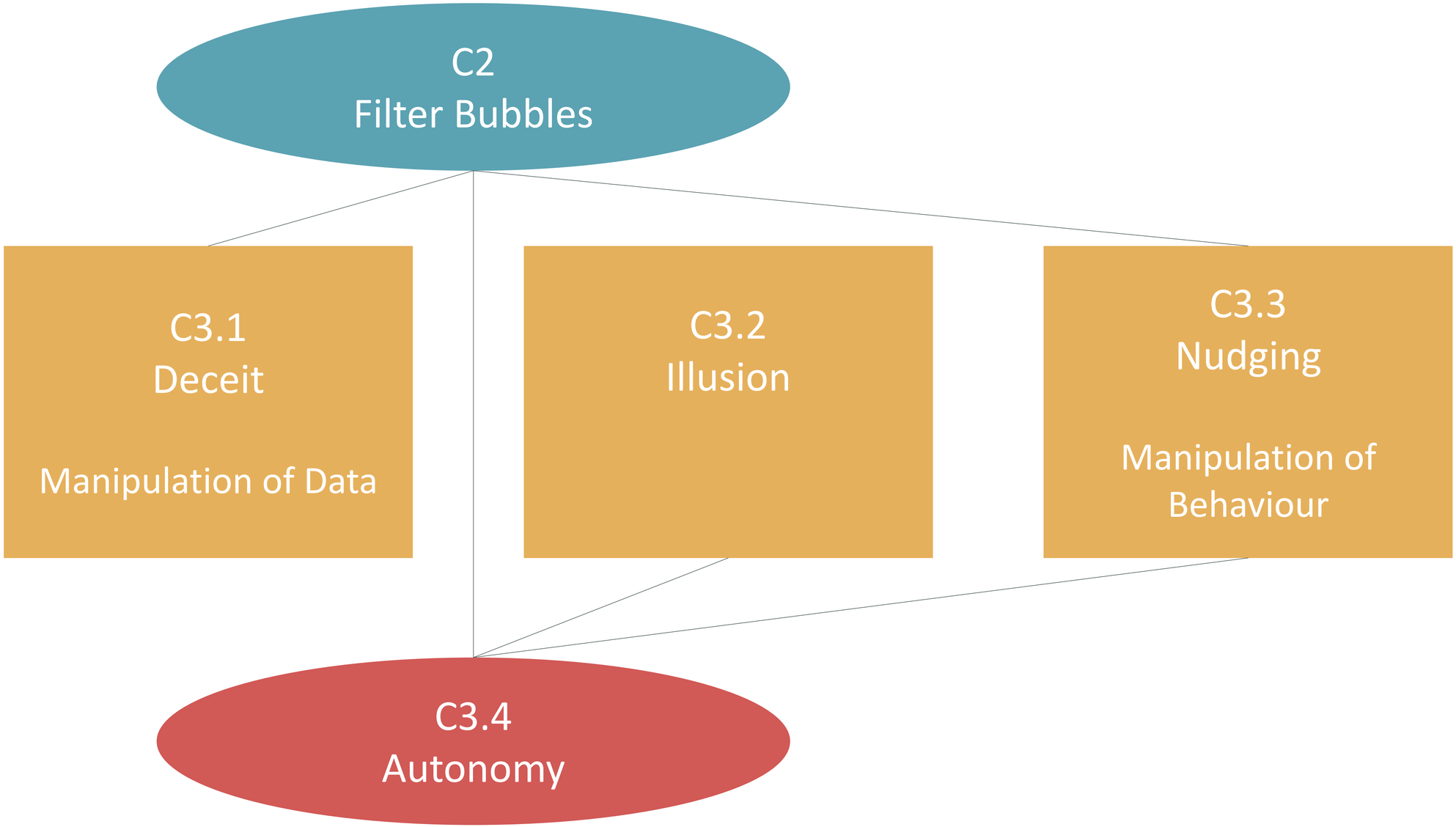
Current Topics C3
Manipulation, Deception, and Illusion



Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

UNIVERSITÄT
DES
SAARLANDES

ROADMAP

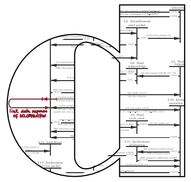

C2
Filter Bubbles
C3.1
Deceit
Manipulation of Data
C3.2
Illusion
C3.3
Nudging
Manipulation of
Behaviour
C3.4
Autonomy

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C3.1
Manipulation, Deception, and Illusion

Deception (Belief-Manipulation)

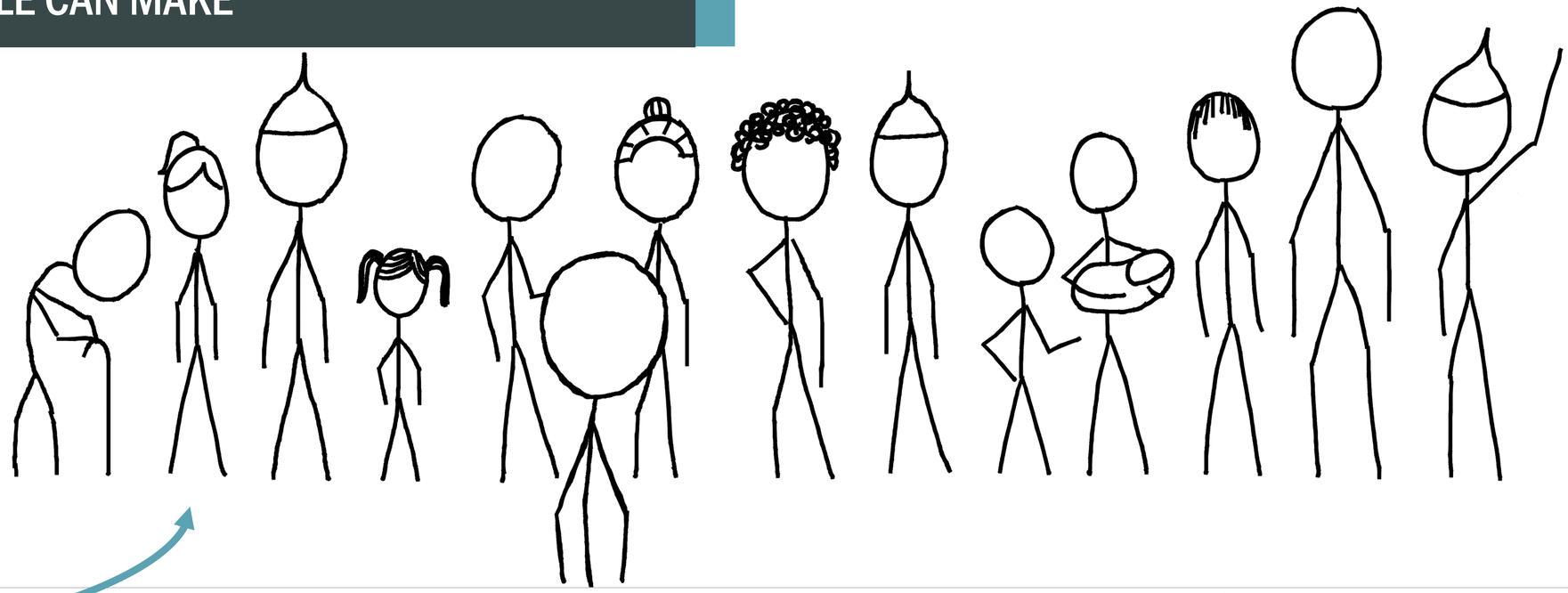


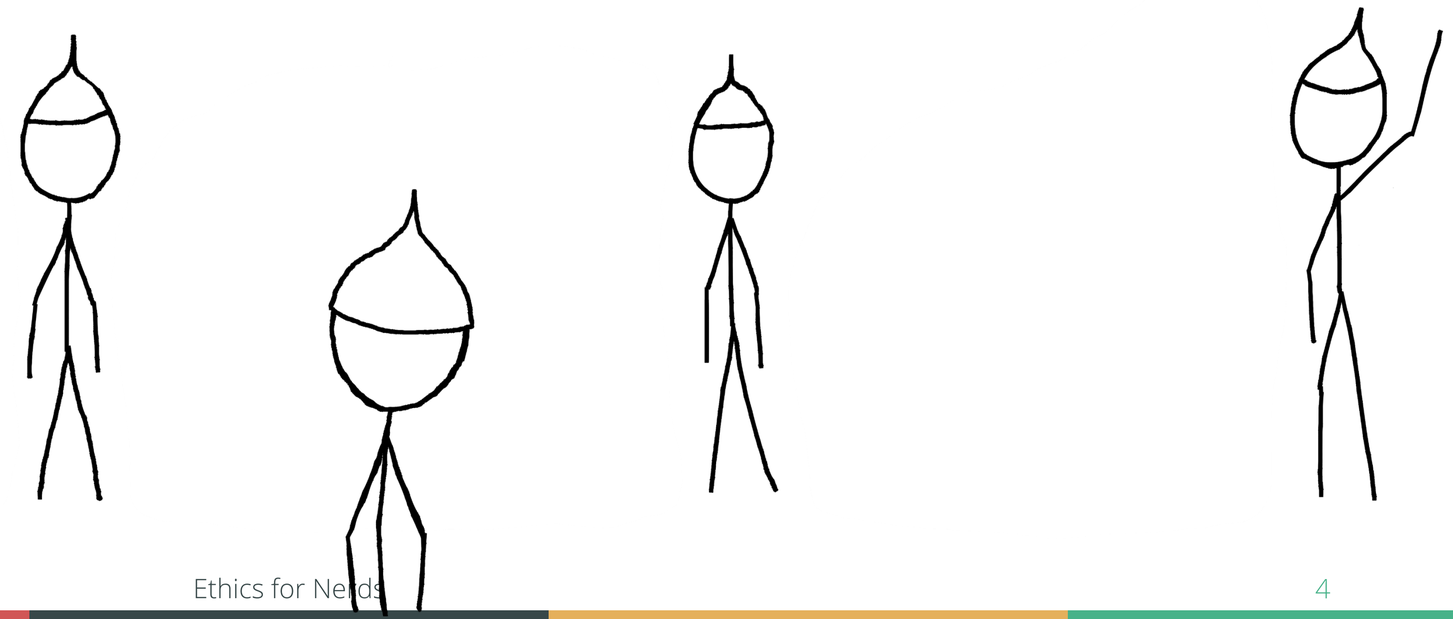
Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

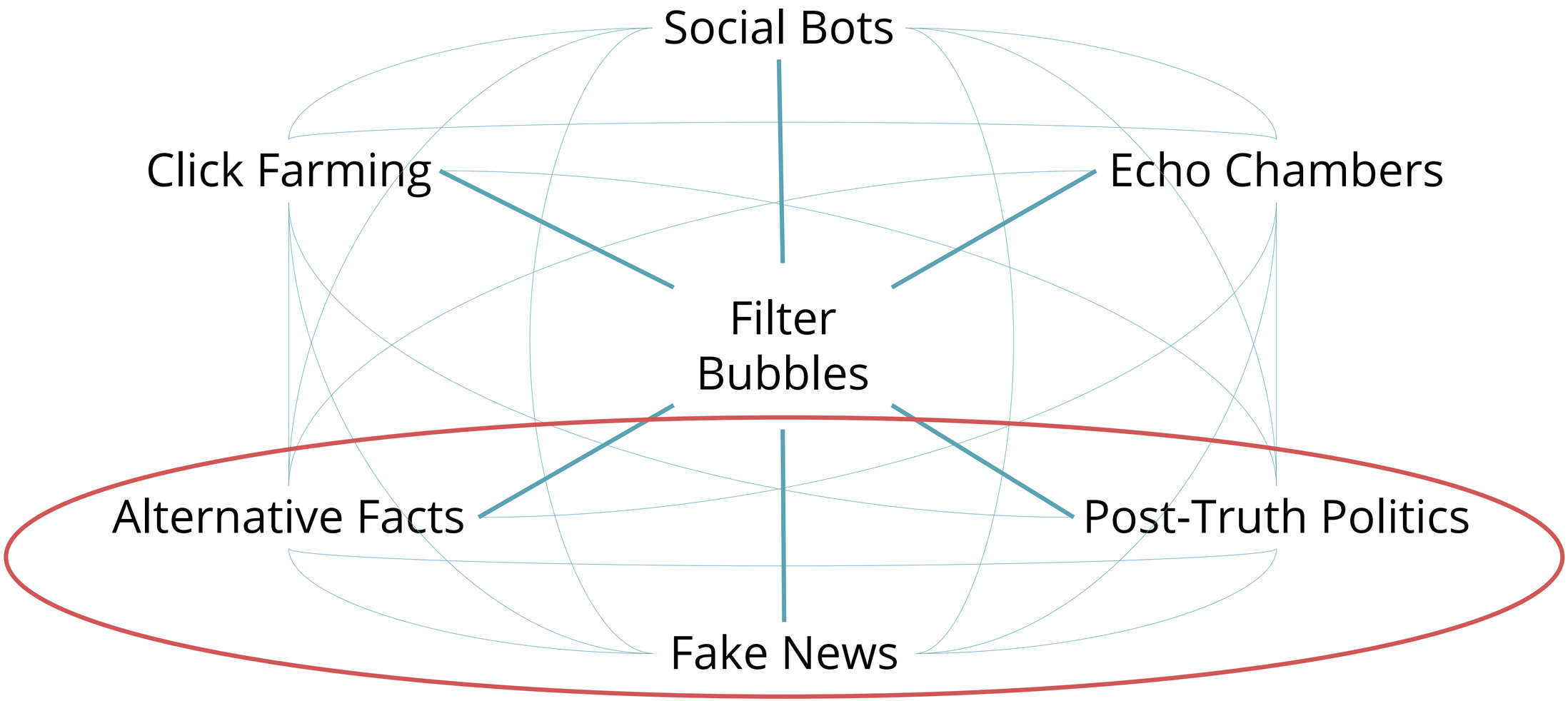
UNIVERSITÄT
DES
SAARLANDES

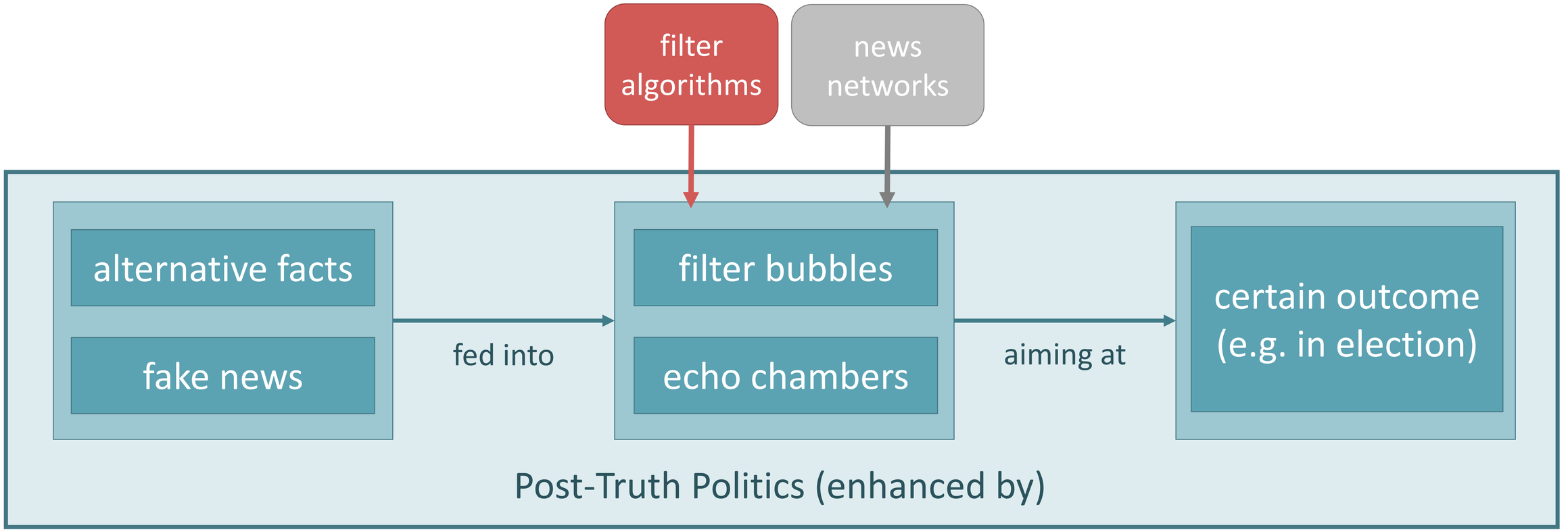
THE DIFFERENCE A FILTER BUBBLE CAN MAKE

Last time we saw:
A filter bubble can make the difference between this and this






Social Bots
Click Farming
Echo Chambers
Filter
Bubbles
Alternative Facts
Post-Truth Politics
Fake News


filter algorithms
news networks
alternative facts
fake news
fed into
filter bubbles
echo chambers
aiming at
certain outcome (e.g. in election)
Post-Truth Politics (enhanced by)

ALTERNATIVE FACTS?

POLITICS | With False Claims, Trump Attacks Media on Turnout and Intelligence Rift

Mr. Spicer also said that security measures had been extended farther down the National Mall this year, preventing “hundreds of thousands of people” from viewing the ceremony. But the Secret Service said the measures were largely unchanged this year, and there were few reports of long lines or delays.

At his first news conference, Sean Spicer, the White House press secretary, accused news outlets of intentionally manipulating photographs “to minimize the enormous support” that President Trump had received at his inauguration. By THE ASSOCIATED PRESS. Photo by Doug Mills/The New York Times. Watch in Times Video »

Embed

Commentary about the size of his inauguration crowd made Mr. Trump increasingly angry on Friday, according to several people familiar with his thinking.

Source:https://www.nytimes.com/2017/01/21/us/politics/trump-white-house-briefing-inauguration-crowd-size.html

Full story: https://www.nytimes.com/2017/01/22/us/politics/president-trump-inauguration-crowd-white-house.html

ALTERNATIVE FACTS?

SECTIONS HOME SEARCH

The New York Times

LOG IN

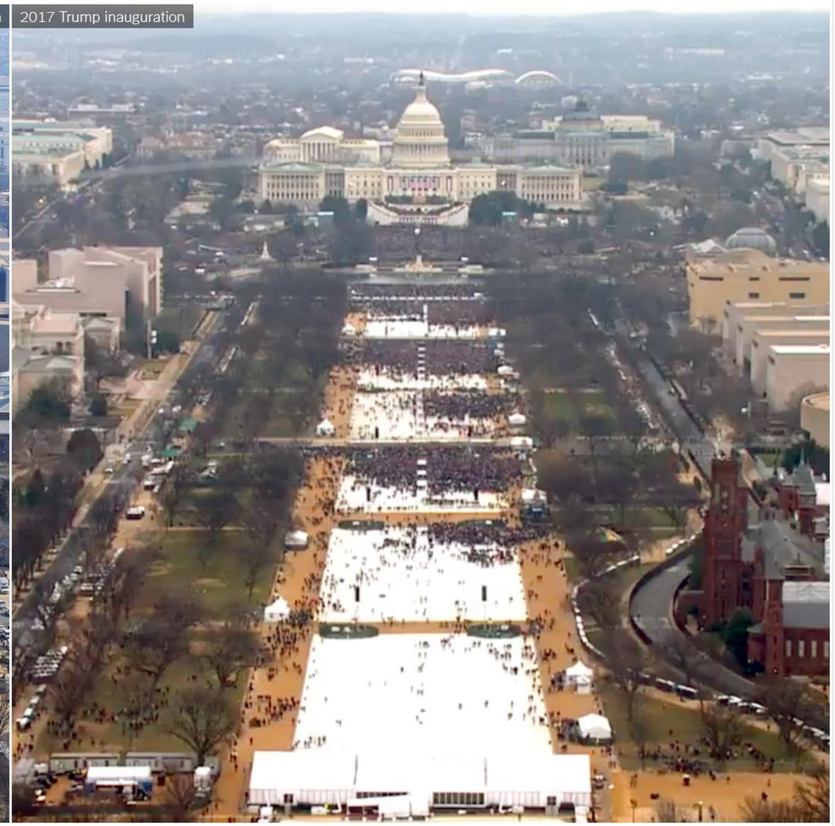
POLITICS

SHARE

Trump's Inauguration vs. Obama's: Comparing the Crowds

By TIM WALLACE, KAREN YOURISH and TROY GRIGGS JAN. 20, 2017

Jewel Samad/AFP/Getty Images

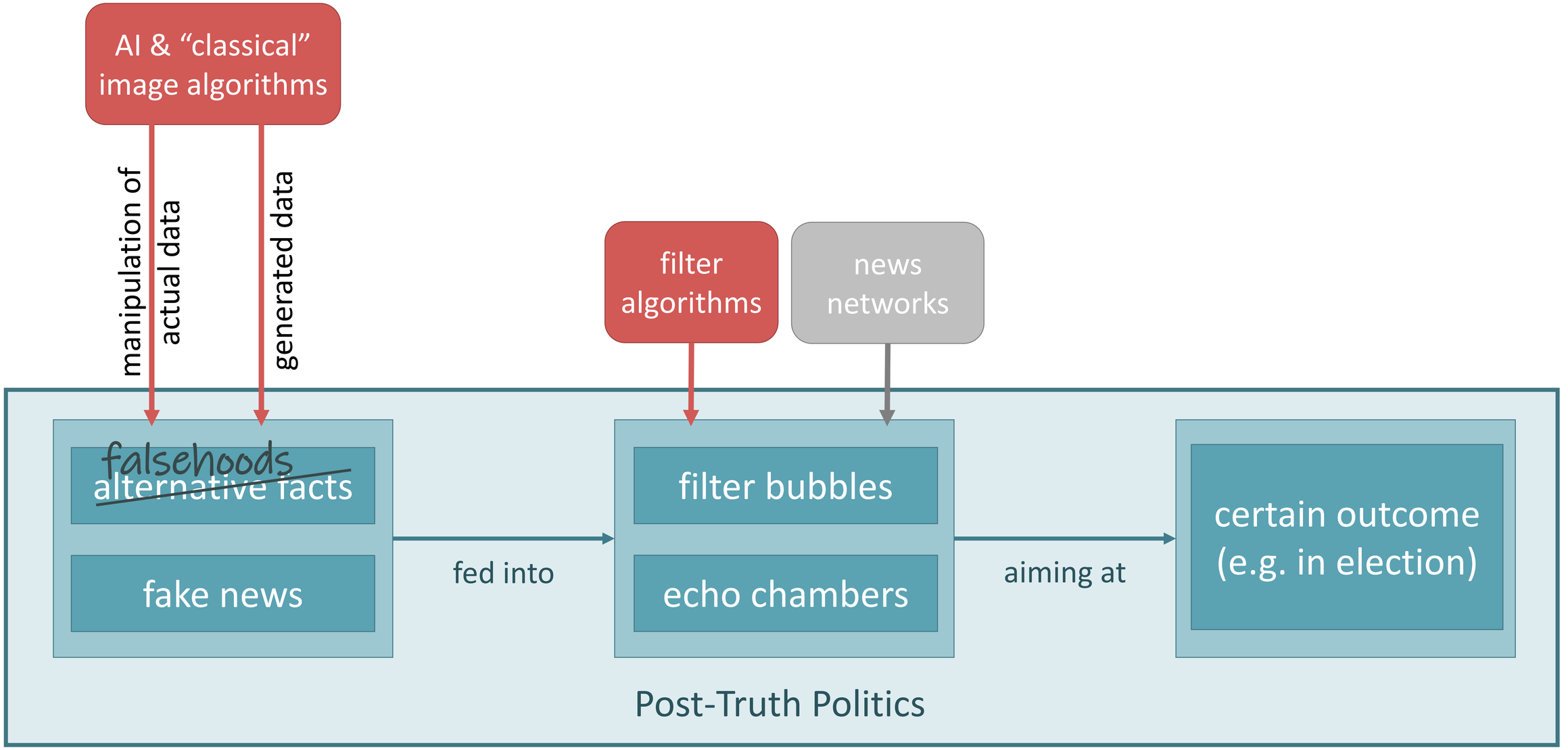


58th Presidential Inaugural Committee

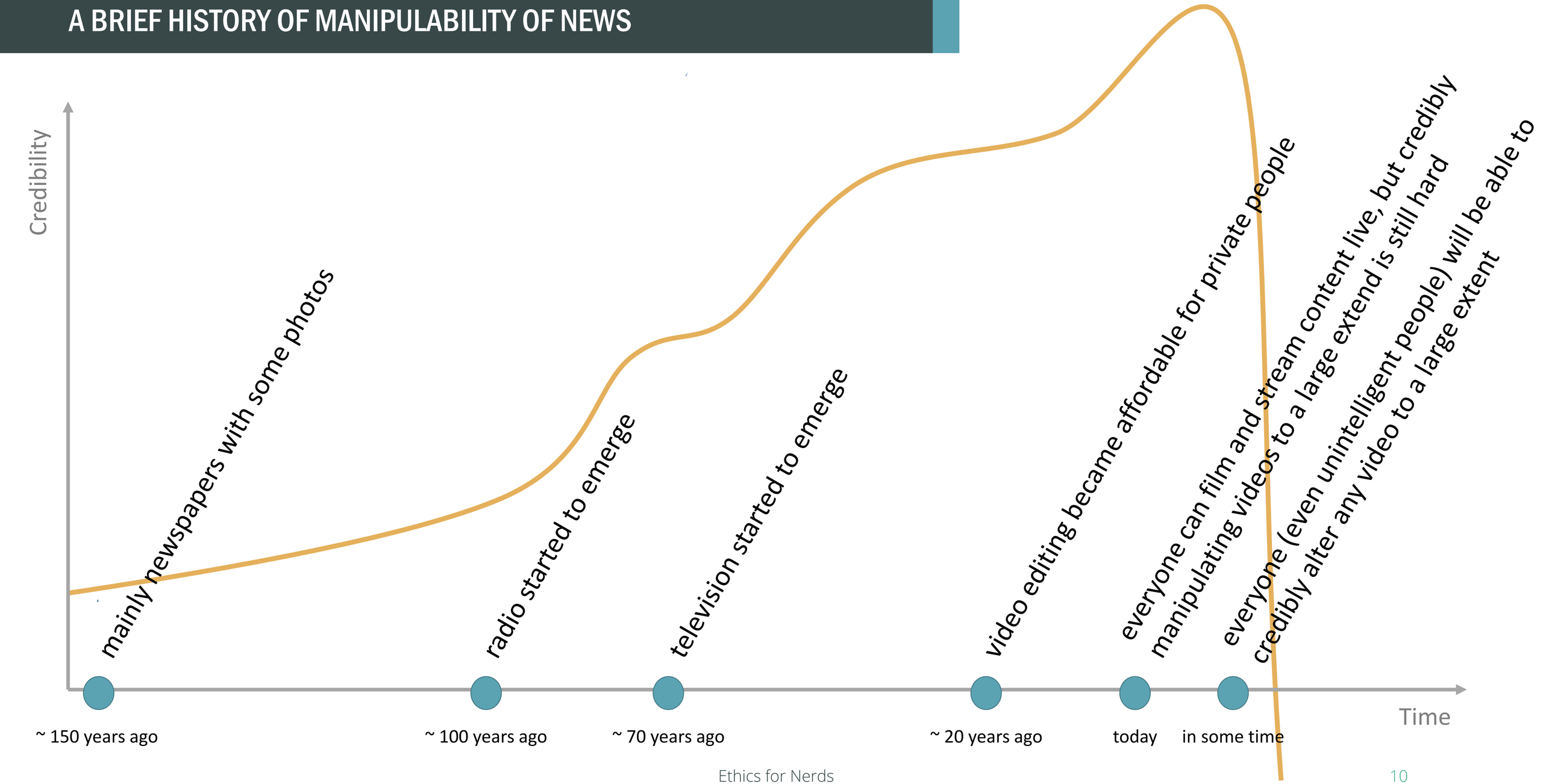
Source: https://www.nytimes.com/interactive/2017/01/20/us/politics/trump-inauguration-crowd.html

Kellyanne Conway, Councler of the US President, coined the phrase "Alternative Facts"

Source:https://en.wikipedia.org/wiki/Alternative_facts


AI & “classical” image algorithms
manipulation of actual data
generated data
filter algorithms
news networks
falsehoods
alternative facts
fake news
fed into
filter bubbles
echo chambers
aiming at
certain outcome (e.g. in election)
Post-Truth Politics

A BRIEF HISTORY OF MANIPULABILITY OF NEWS



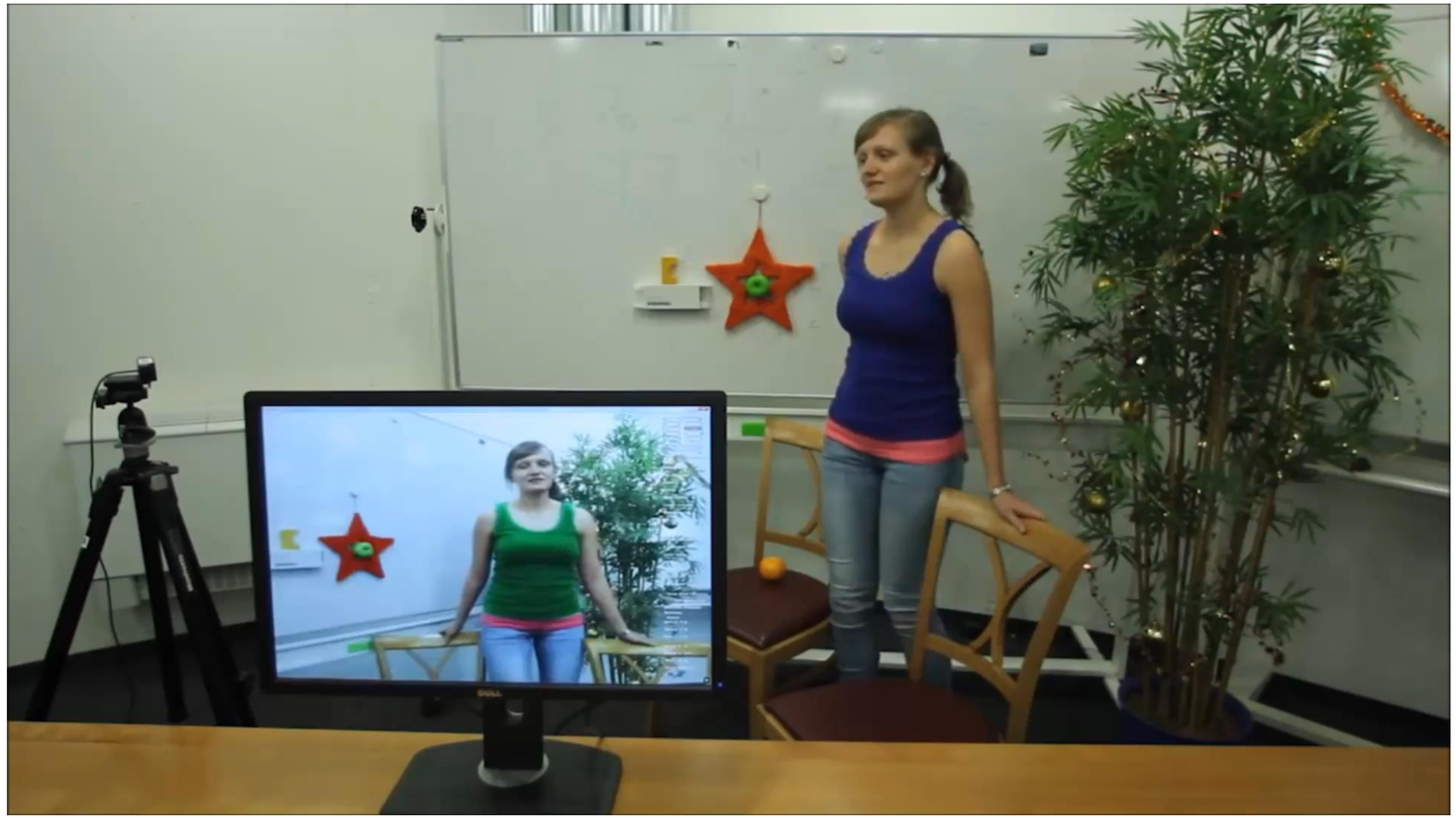
VAROUFAKE - A POPULAR EXAMPLE OF MANIPULATION

Source: https://youtu.be/Vx-1LQu6mAE

SOME TOOLS

Live Intrinsic Video (Meka, Zollhöfer, Richardt, Theobalt)

- alters the color and texture in a video in real time



http://gvv.mpi-inf.mpg.de/projects/MZ/Papers/SG2016_IV/page.html

https://www.bundestag.de/dokumente/textarchiv/2014/-/286976

SOME TOOLS

Adobe Voco

- Software for editing speech on basis of text, what you type is what you hear
- all you need is 20 minutes of (any!) speech of the person you want to edit as training

original sentence

sentence after alteration

SOME TOOLS

Face2Face (Thies, Zollhöfer, Stamminger, Theobalt, Nießner)

http://www.graphics.stanford.edu/~niessner/thies2016face.html

SOME TOOLS

Deep Fake

- free AI system that changes faces in video for different (potential) purposes:

https://www.youtube.com/watch?v=lpk7ocOc2ho

post-production of movies

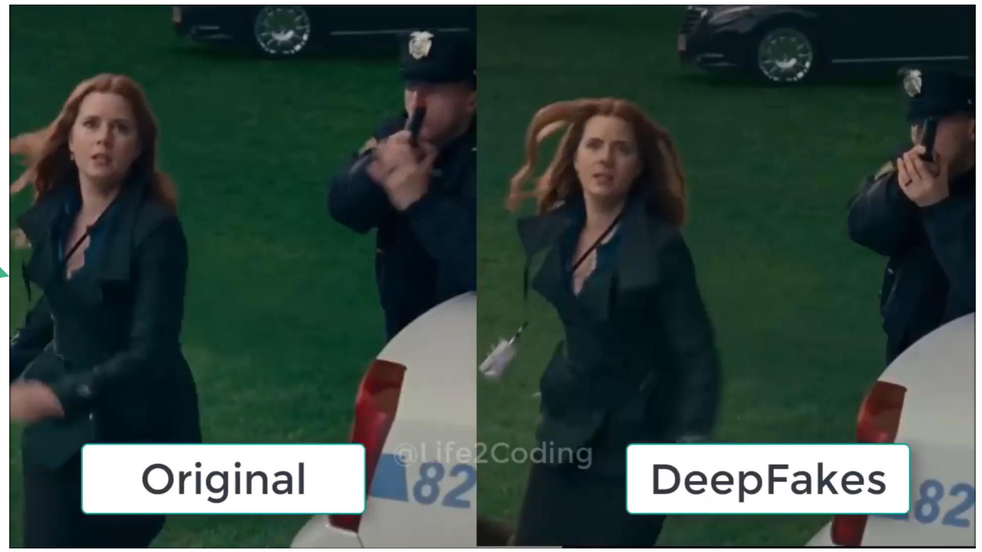
producing fake news

https://www.youtube.com/watch?v=knRGxj37AjM

comedy

porn

[omitted]



https://www.youtube.com/watch?v=knRGxj37AjM

SOME TOOLS

Technology like this is great, in many ways:

- for filmmakers (films with dead actors, lowers the costs of postproduction)
- could be used to make better augmented reality applications, e.g. for industry or disabled people
- and it’s an interesting technology in itself (isn’t it cool what’s possible?)

And technology like this is not so great, in many ways, too:

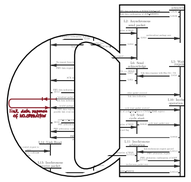
- first of all: **Misuse!**
- credibility of news weakened (even if not manipulated)
- easier for conspiracy theorists and extremists to muffle themselves and others in cozy immunized echo chambers
- lots of people do not have the technical know-how yet to tell what could likely be faked and what couldn’t
- technical problems, like bypassing biometrical security systems

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C3.2
Manipulation, Deception, and Illusion

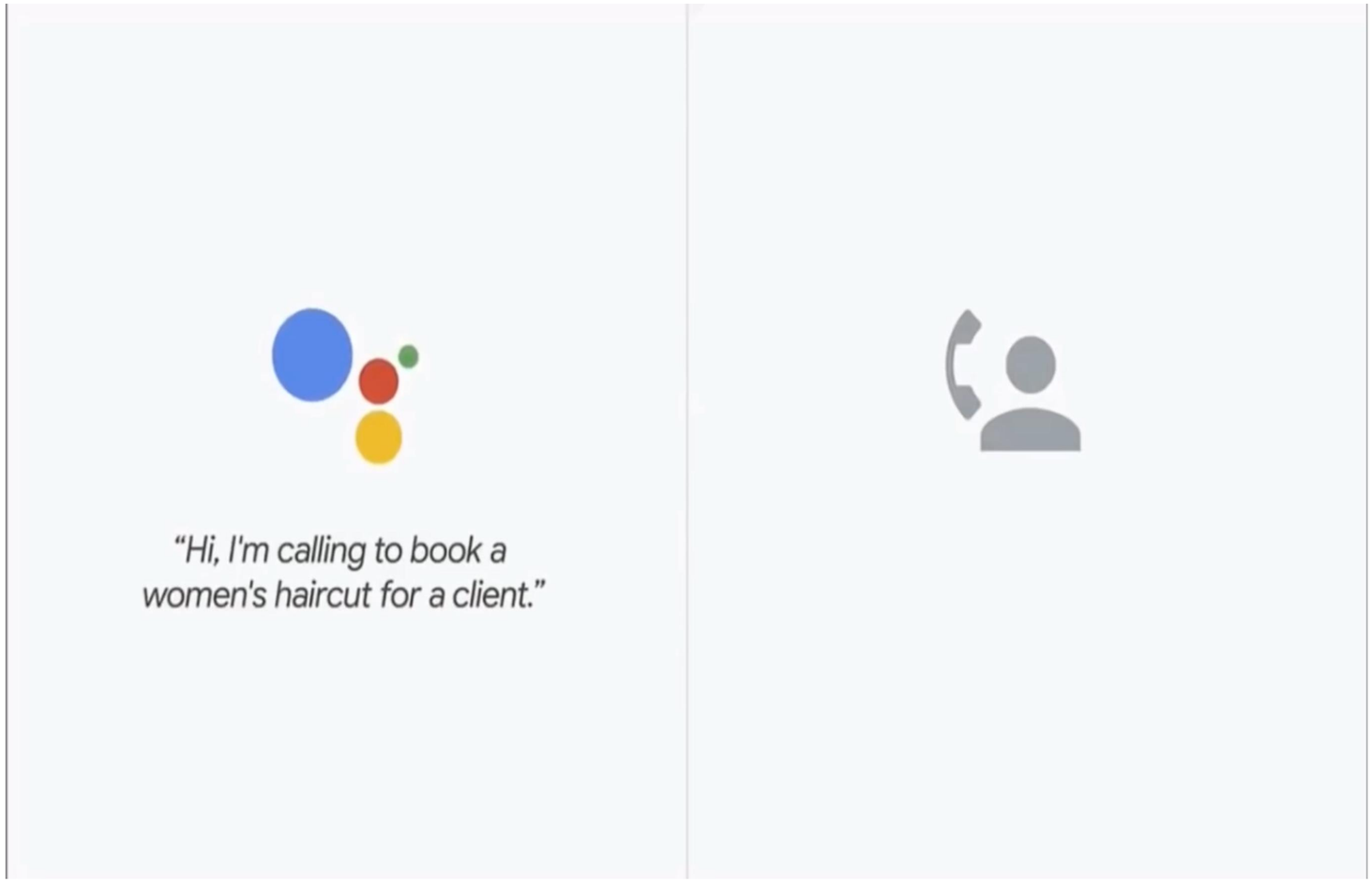
Illusion and Other Phenomena



Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

Google Duplex

- AI system that automatically makes calls for making an appointment or a reservation
- not distinguishable from human if conversation goes as planned



https://www.youtube.com/watch?v=bd1mEm2Fy08

RED FLAG LAW

Red Flag Act

originally was a law in the UK that said that in front of every motorized vehicle there has to be someone walking with a red flag in order to warn pedestrians

https://dailybritain.wordpress.com/tag/red-flag-act/

RED FLAG LAW

The most infamous of the Red Flag Laws was enacted in Pennsylvania circa 1896, when legislators unanimously passed a bill through both houses of the state legislature, which would require all motorists piloting their "horseless carriages", upon chance encounters with cattle or livestock to (1) immediately stop the vehicle, (2) "immediately and as rapidly as possible ... disassemble the automobile", and (3) "conceal the various components out of sight, behind nearby bushes" until equestrian or livestock is sufficiently pacified. The bill did not become law, as the Governor of Pennsylvania used an executive veto.

https://en.wikipedia.org/wiki/Red_flag_traffic_laws

Red Flags for AI

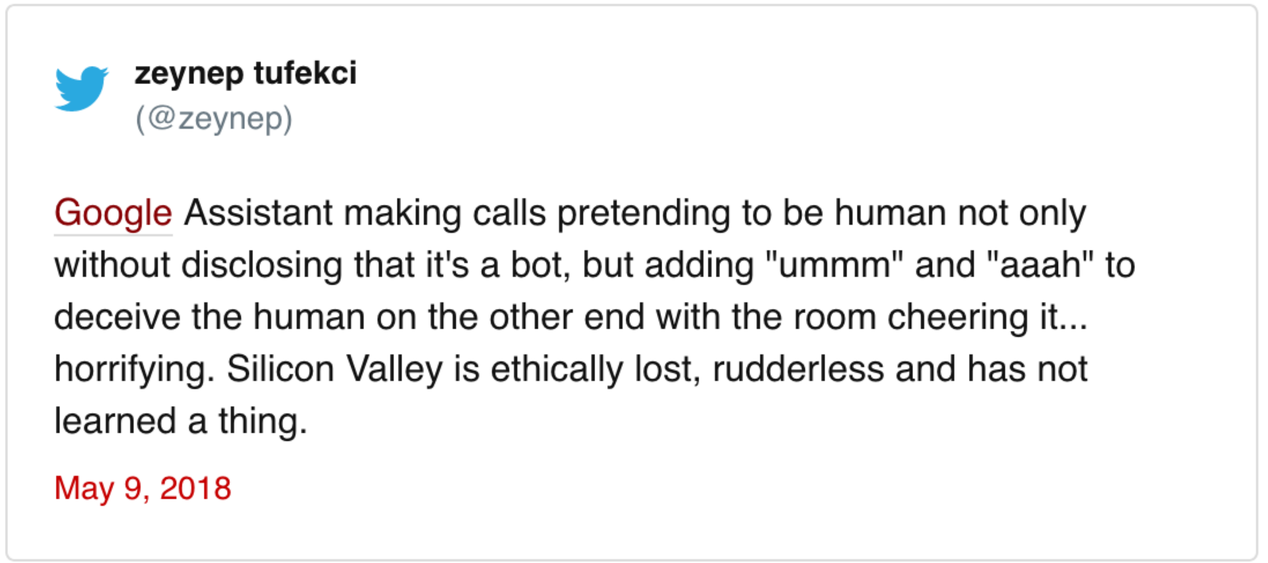
whenever an AI interacts with a human, it has to be made explicit to the human that they are interacting with an AI

Advantage:

- humans cannot mix up AIs and other humans

Disadvantage:

- almost none (?)



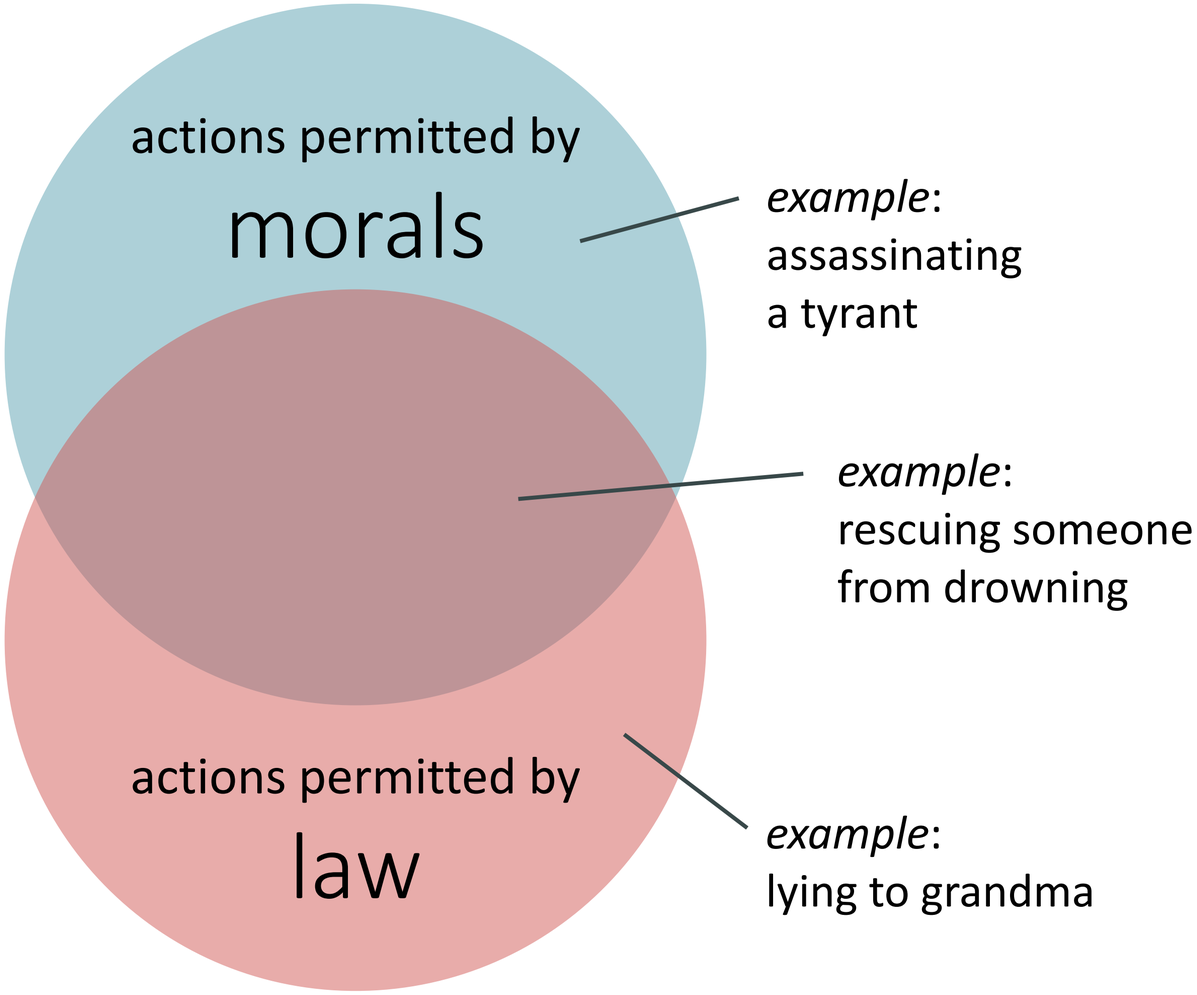
https://www.theguardian.com/technology/2018/may/11/google-duplex-ai-identify-itself-as-robot-during-calls

Morals

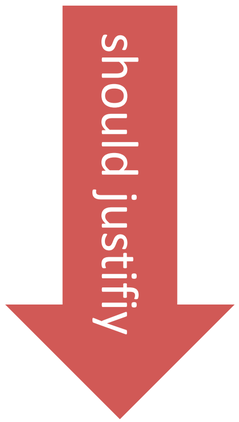
should justifiy

Law

but still
(even if some
laws are justified)



Morals



Red Flag Law

Why it could be morally obligatory to introduce a Red Flag Law

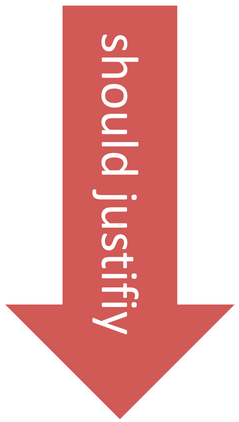
both Utilitarianism and Scanlon agree that it is obligatory to introduce a Red Flag Law:

Utilitarianism (easy)

- introduction is not costly,
- but clearly has many benefits

→ overall introducing a Red Flag Law most likely is the option with the overall best utility

Morals



Red Flag Law

Why it could be morally obligatory to introduce a Red Flag Law

both Utilitarianism and Scanlon agree that it is obligatory to introduce a Red Flag Law:

Scanlon (very briefly)

- current AI is not general AI and thus cannot adequately react to all situations, especially non-standard situations
- if there is an emergency, the AI cannot react appropriately
- a human who thinks to interact with another human in an emergency, but is really talking to an AI, can reasonably reject all sets of principles that do not demand a red flag law

→ not introducing a Red Flag Law is wrong, i.e. introducing a Red Flag law is right

Paro

a therapeutic robot

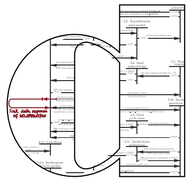
Image and more information: https://robots.ieee.org/robots/paro/
and https://www.wohlfahrtswerk.de/innovation-und-projekte/praxis-und-innovationsprojekte/paro/ (video has English subtitles)

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

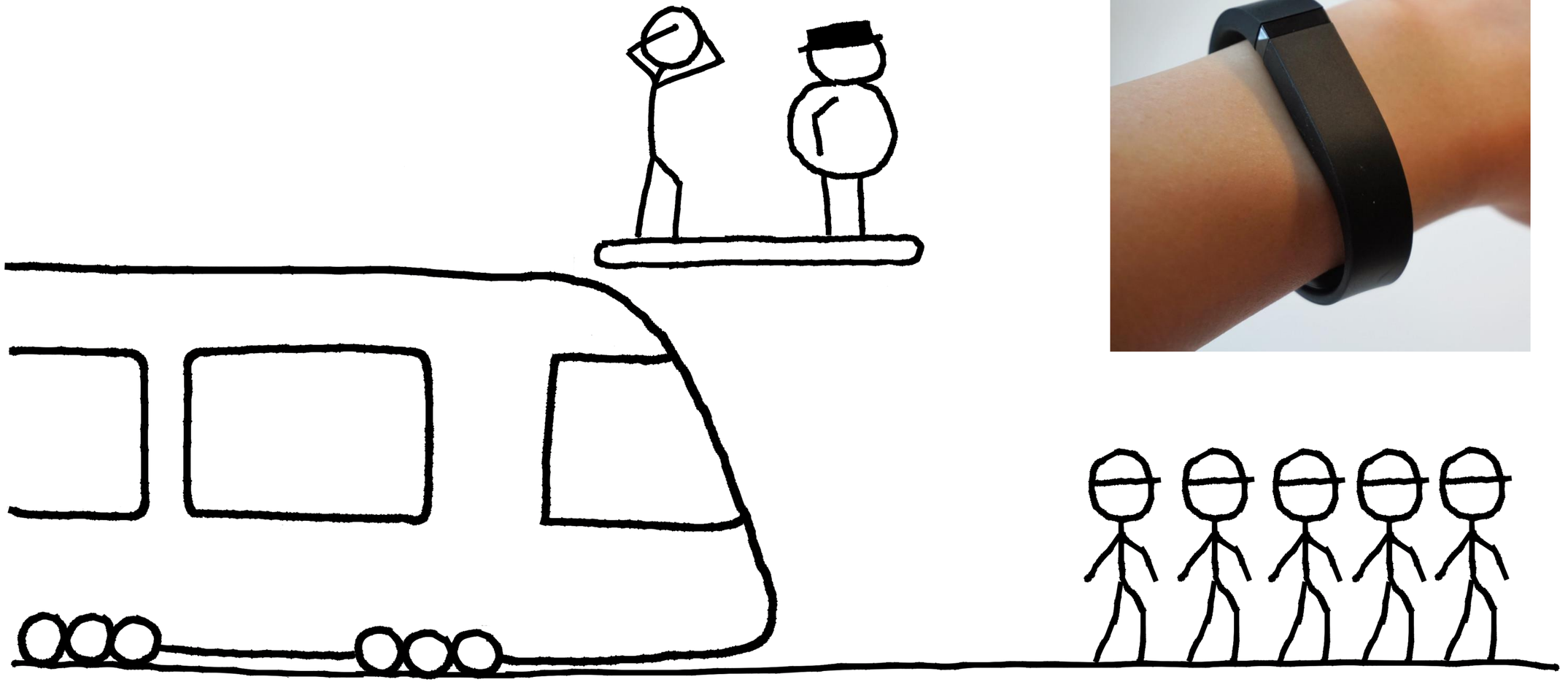
Current Topics C3.3
Manipulation, Deception, and Illusion

Nudging (Behaviour-Manipulation)



Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

FAT MAN



FAT MAN

P1	Fitness trackers can contribute to increasing one’s activity.
P2	Increasing one’s activity benefits one’s health.
P3	Improving our health usually is beneficial for us.
P4	If improving our health usually is beneficial for us and fitness trackers can contribute to benefitting one’s health, then fitness trackers are a good thing.
C	Therefore, fitness trackers are a good thing.

Or aren’t they?

(Leaving privacy issues aside.)

nudging

influencing the behaviour of people without establishing new rules or changing economical incentives

https://commons.wikimedia.org/wiki/File:Urinal_Fly.JPG?uselang=de

https://www.supermarktblog.com/2015/04/16/sonderangebote-verstehen-in-nur-3-minuten/

- having an especially overpriced item on the menu
- opt-out instead of opt-in
- …

PERSUASIVE TECHNOLOGY

These usually are bad. Period.

persuasive technology

technology that tries to persuade its user to adopt a new behaviour, but not in a coercive way

gamification

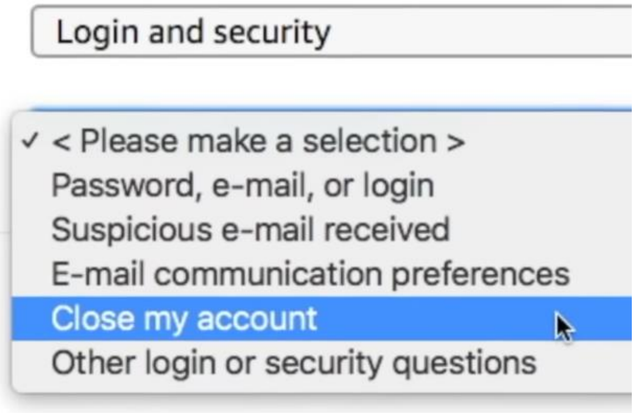
the use of game elements in non-game contexts

dark pattern

UI design that is made to trick users into doing things that may not be in their interest

...

https://www.game.de/publikationen/videospiele-im-alter-warum-immer-mehr-senioren-spielen/



Nerdwriter1: *How Dark Patterns Trick You Online*
https://youtu.be/kxkrdLI6e6M

For more examples, look here:
https://www.darkpatterns.org/

Possible Upsides of Digital Nudging

helping people to be

- more motivated
- perform tasks they do not like much easier
- get/stay healthy
- adopt behaviours that are beneficial for them or for the society

https://www.game.de/publikationen/videospiele-im-alter-warum-immer-mehr-senioren-spielen/

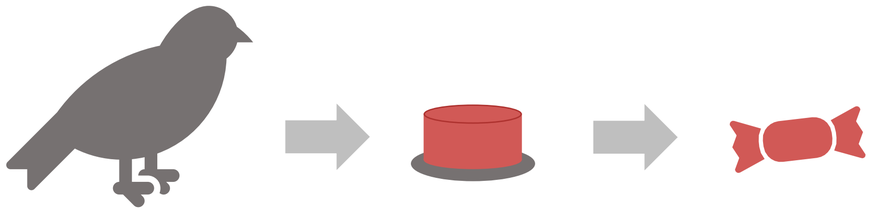
SLOT MACHINES

https://commons.wikimedia.org/wiki/File:Las_Vegas_slot_machines.jpg

https://www.flickr.com/photos/kalleboo/4652948194

SLOT MACHINES

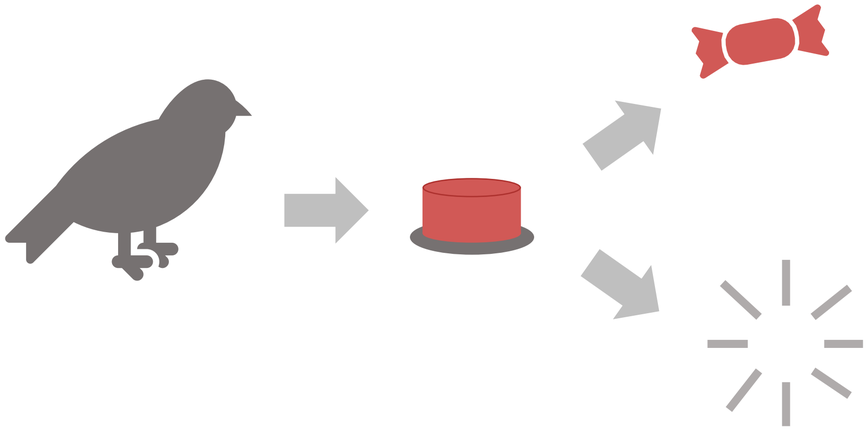
Skinner's Box



subject gets rewarded every time the button is pushed

↓

subject will get bored and stop pushing the button



subject gets rewarded only sometimes when the button is pushed

↓

subject pushes button almost endlessly

something about not knowing what to expect is very exiting for the subject

SLOT MACHINES

SLOT MACHINES

Schüll, N. D. (2014). *Addiction by design: Machine gambling in Las Vegas*. Princeton University Press.

http://www.peakpx.com/528611/casino-interior

low ceiling

mostly indirect light

little echo

http://www.peakpx.com/403839/black-casino-slot-machine

no light surfaces that make light bounce of

comfy chairs

no right angles

modern casinos are designed in a way such that nothing makes you stop and think

"continuous gaming productivity"

SLOT MACHINES

the casino wants

money

the gambler wants

a certain experience

GOING BEYOND CASINOS

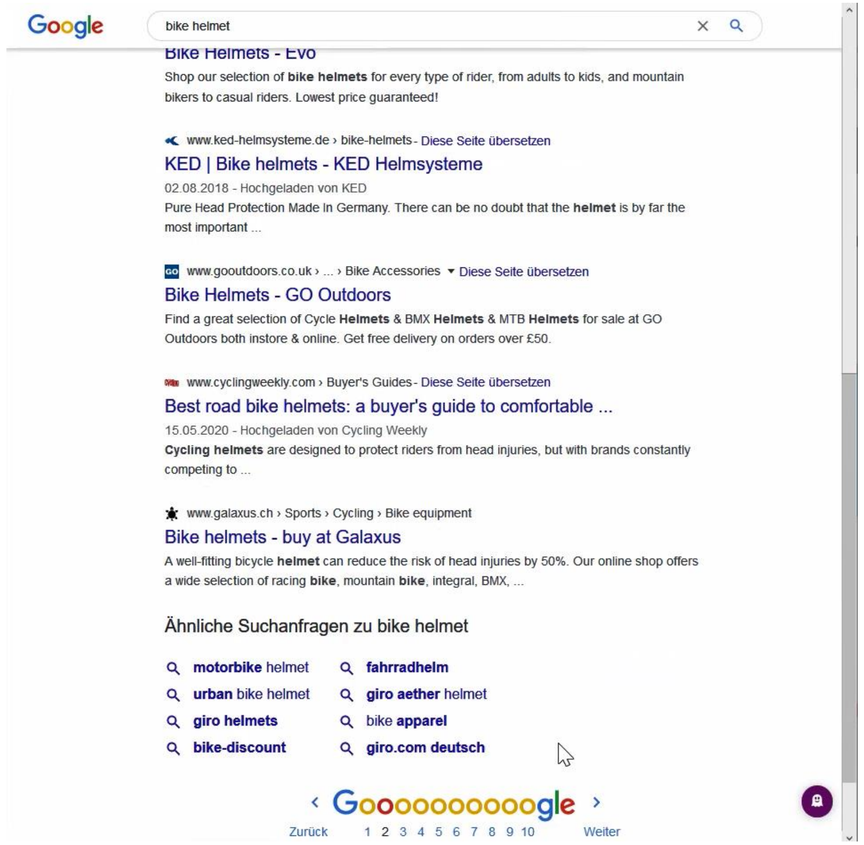
the provider wants

money

the user wants

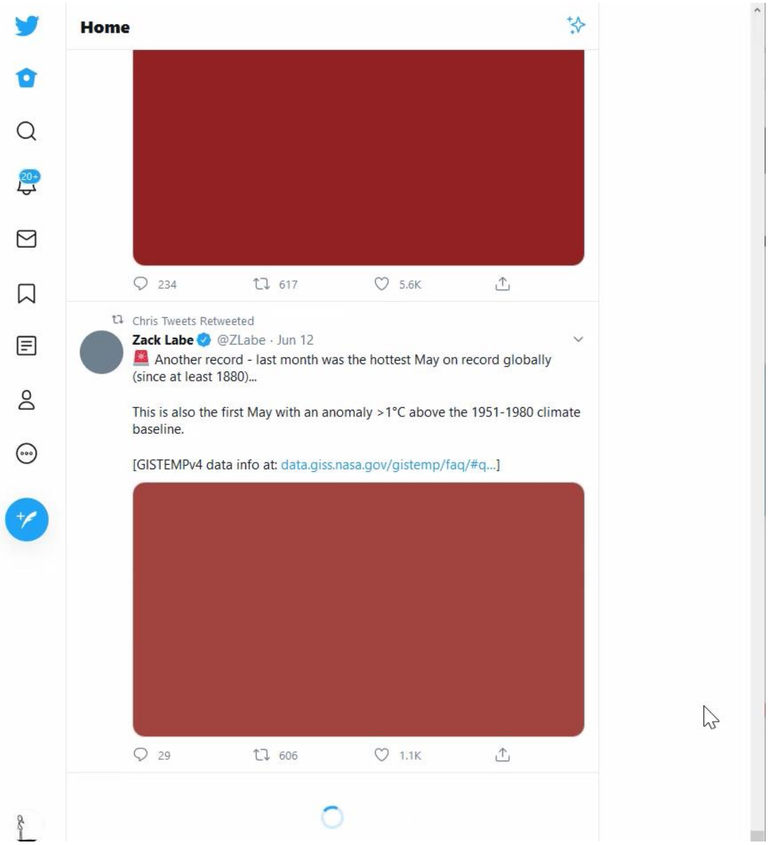
a certain experience

GOING BEYOND CASINOS



pagination

you reach the bottom of the page rather quickly when you scroll down



infinite scroll

new content is loaded dynamically while you scroll down

GOING BEYOND CASINOS

Lots of things you see in technology are related to mechanics of slot machines

infinite scroll

no right angles

GOING BEYOND CASINOS

Lots of things you see in technology are related to mechanics of slot machines

phone buzzes

you check it

maybe it gives you an exciting notification

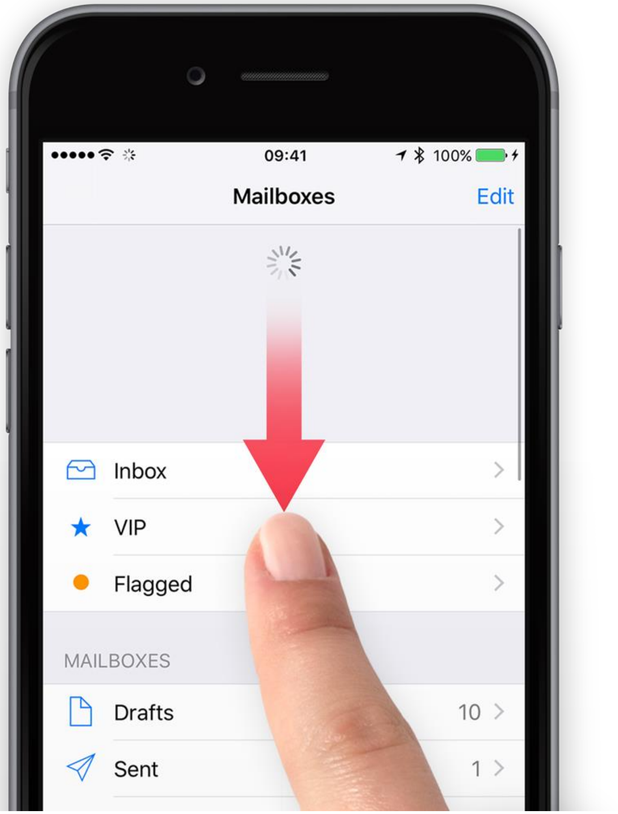
you push a button

maybe the machine gives you money back

GOING BEYOND CASINOS

Lots of things you see in technology are related to mechanics of slot machines

http://www.tapsmart.com/tips-and-tricks/refresh-inbox-pull-refresh-emails-ios-12-guide/



you pull down your feed to refresh

maybe you will see an exciting post

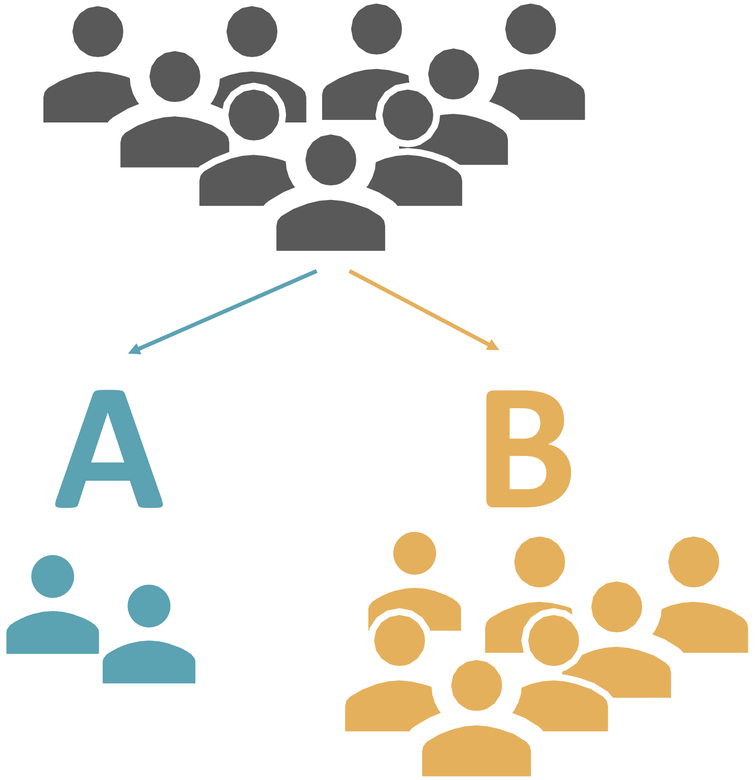
you push a button

maybe the machine gives you money back

Your ndivided attention

Side note: A-B-tests

instead



some mechanisms are very hard to avoid, because they utilizes certain ways are brains work

Hooking up to goal-less games

Coin Master

Candy Crush Saga

Idle Miner Tycoon

Hooking up to goal-less games

slot-machine like mechanics, like loot boxes, or even in-game slot machines

addiction

you loose money

no real satisfaction

https://vimeo.com/154271693

PROS OR CONS?


Überfalle das Dorf
eine Freundes


Greife die Dörfer
deiner Freunde an !

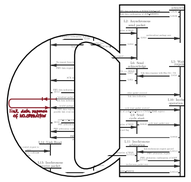

Coin Master
Bjorn
434 000
JACKPOT!
30/50
SPIN
Knacke den
Jackpot !

Ethics for Nerds

An Advanced Course in Computer Science
Summer Semester 2020

Current Topics C3.4
Manipulation, Deception, and Illusion

Autonomy



Prof. Holger Hermanns,
Kevin Baum, Sarah Sterz

AUTONOMY

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

What is autonomy?

Intuition: you are autonomous if you are not guided by outside forces too much

Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

What mental states?

highest-order desires regarding which of her first-order desires moves her to act

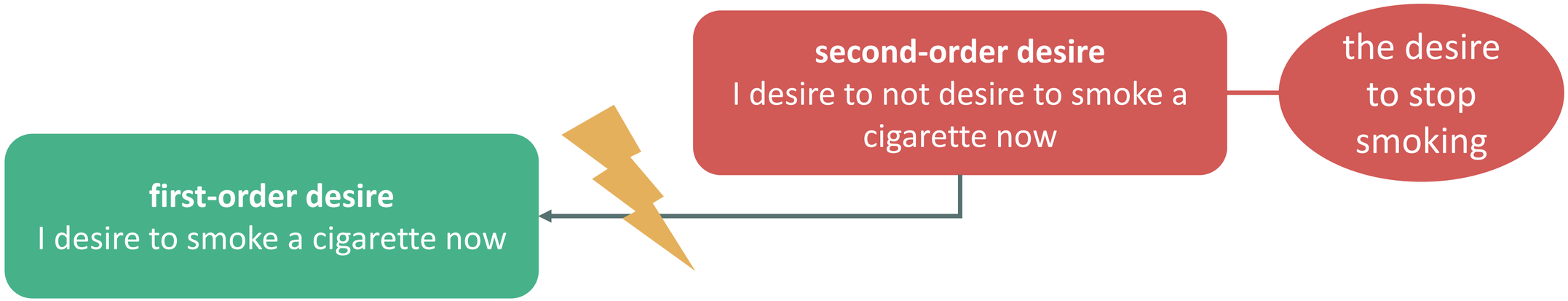
Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

Higher-order desires

a higher-order desire is a desire that is about another desire



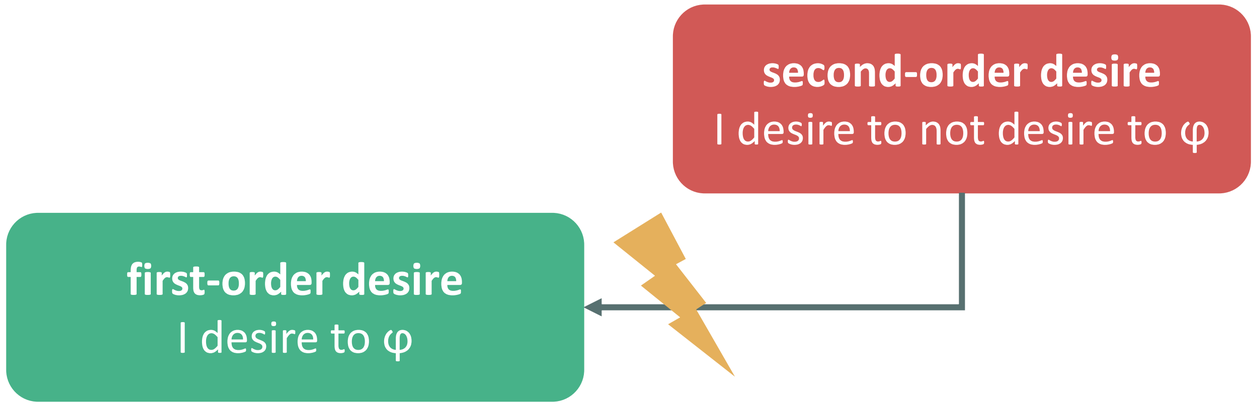
Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

What mental states?

highest-order desires regarding which of her first-order desires moves her to act



agent φs: agent does not govern her own action

agent does not φ: agent does govern her own action

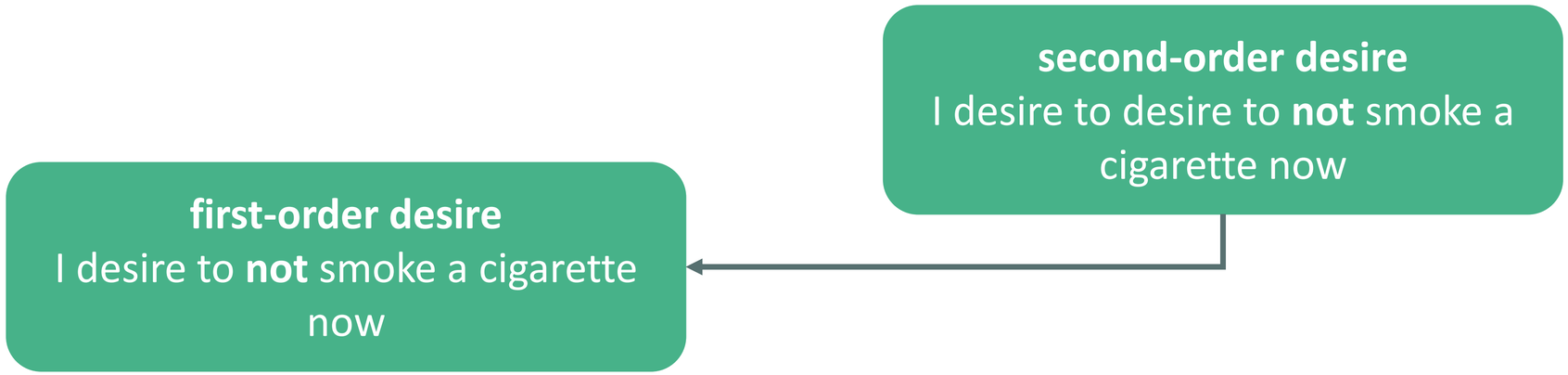
Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

Higher-order desires

a higher-order desire is a desire that is about another desire



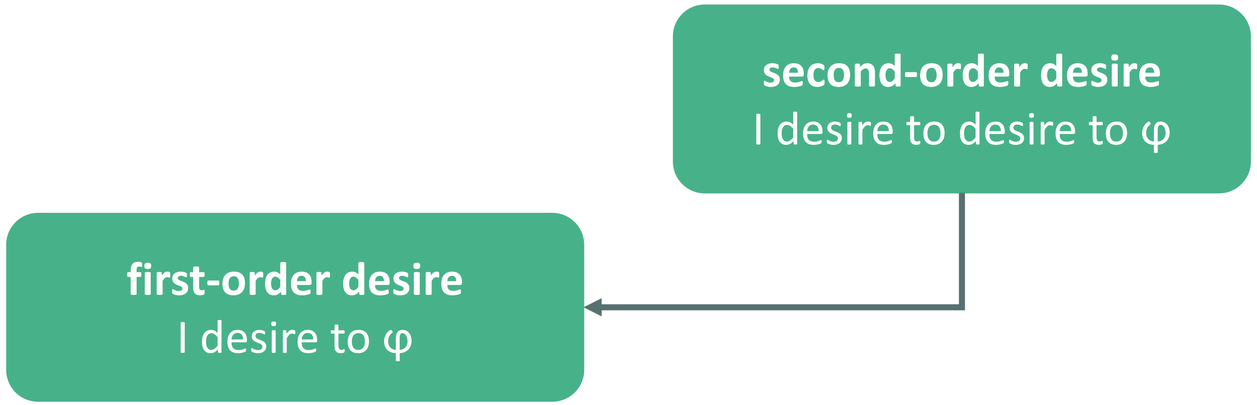
Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

What mental states?

highest-order desires regarding which of her first-order desires moves her to act



agent φs: agent does govern her own action

agent does not φ: agent does not govern her own action

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

What mental states?

- highest-order desires regarding which of her first-order desires moves her to act
- judgments regarding which actions are (most) worth performing
- additionally there must be harmony between what the agent does and her long-term plans
- relatively stable network of emotional states constitutive of “caring”
- agent’s character traits
- most thoroughly “integrated” psychological states

AUTONOMY

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

What is autonomy?

Intuition: you are autonomous if you are not guided by outside forces too much

Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

Action Government – Reasons-Responsiveness View

An agent governs her own actions only if her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does.

AUTONOMY

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

What is autonomy?

Intuition: you are autonomous if you are not guided by outside forces too much

Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

Action Government – Incompatibilist View

An agent governs her own actions only if her actions cannot be fully explained as the effects of causal powers that are independent of her, even if her beliefs and attitudes are among these effects.

Laplace's Demon and Determinism

https://de.wikipedia.org/wiki/Laplacescher_D%C3%A4mon

We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.

Pierre Simon Laplace, A Philosophical Essay on Probabilities (1814)

INTERLUDE: FREEDOM

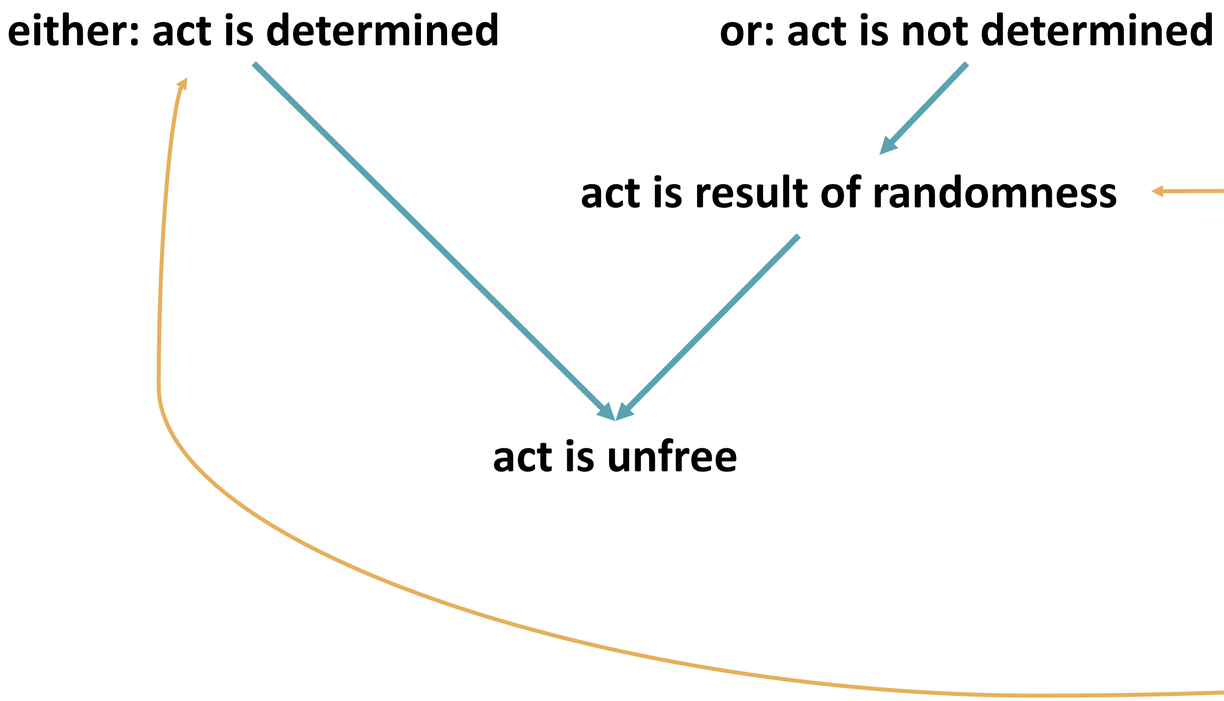
determined

not determined

INTERLUDE: FREEDOM

(partly adapted from slides of the lecture *Introduction to Practical Philosophy* by Prof. Christoph Fehige)

The Either-Or-Problem of Freedom



Ways to Get Autonomy In

Possible alternative to randomness(?):
Allow also for some kind of *autonomy*

Determined, but by the right thing

- **Coherentist View**: the right mental states
- **Reasons-Responsiveness View**: the right kind of reasons played a sufficiently large role in the process

DOES NUDGING TAKE AWAY OUR AUTONOMY?

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

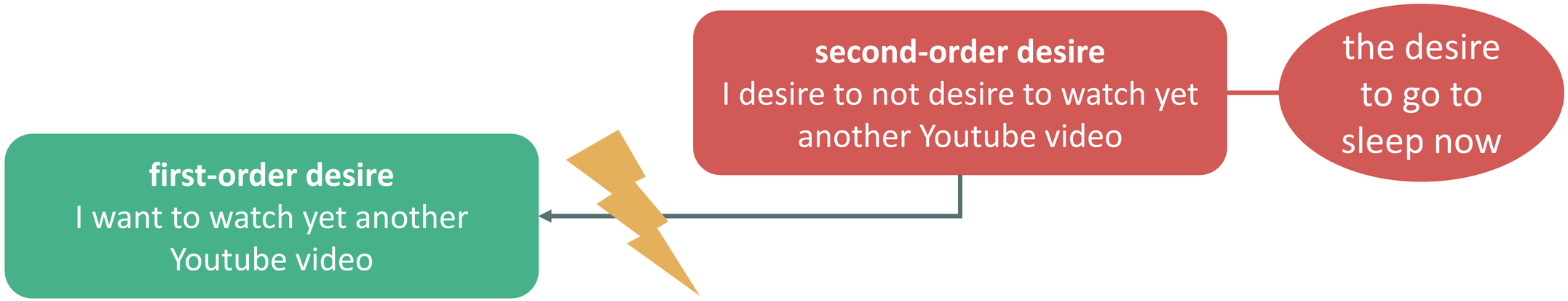
Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

Yes, if it 'overrides' our relevant metal states in the process



DOES NUDGING TAKE AWAY OUR AUTONOMY?

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

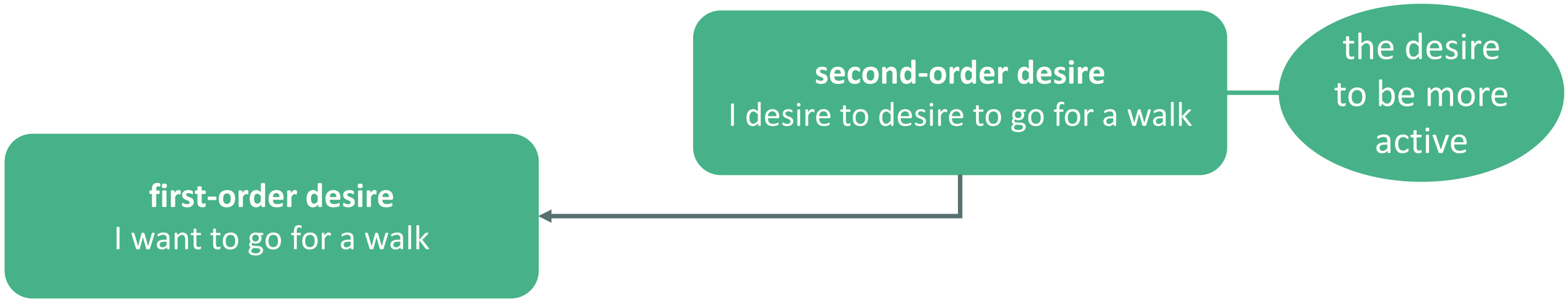
Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

Action Government – Coherentist View

An agent governs her own action if and only if she is motivated to act as she does because this motivation coheres with some mental state that represents her point of view on the action.

No, if it does not ‘override’ our relevant metal states in the process



DOES NUDGING TAKE AWAY OUR AUTONOMY?

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

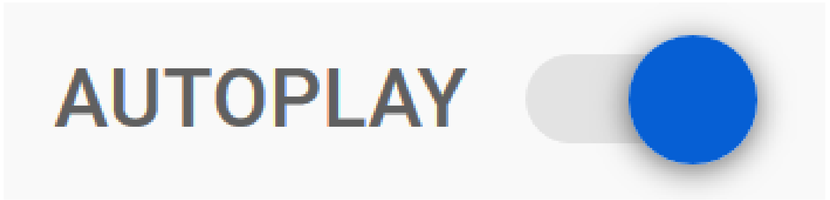
Autonomy (working definition)

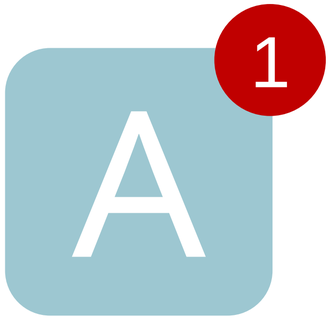
An agent is autonomous if and only if she governs her own action.

Action Government – Reasons-Responsiveness View

An agent governs her own actions only if her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does.

Yes, if it makes us sufficiently reason-unresponsive





DOES NUDGING TAKE AWAY OUR AUTONOMY?

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

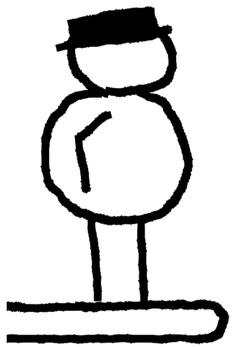
Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

Action Government – Reasons-Responsiveness View

An agent governs her own actions only if her motives, or the mental processes that produce them, are responsive to a sufficiently wide range of reasons for and against behaving as she does.

No, if it does not makes us sufficiently reason-unresponsive



DOES NUDGING TAKE AWAY OUR AUTONOMY?

Buss, Sarah and Westlund, Andrea, "Personal Autonomy", *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/

Autonomy (working definition)

An agent is autonomous if and only if she governs her own action.

Action Government – Incompatibilist View

An agent governs her actions only if her actions cannot be fully explained as the effects of causal powers that are independent of us, even if our beliefs and attitudes are among these effects.

we cannot tell just by this definition of autonomy

DOES NUDGING TAKE AWAY OUR AUTONOMY?

It is very plausible that some nudging technologies can take away our autonomy.

Is that bad?

Depends on whether autonomy is valuable!

Is it?

Depends on axiology!

- intrinsically
- extrinsically

IS THAT A PROBLEM?

Is autonomy intrinsically valuable?

Hedonism	Preference Theory	Objective List Accounts
pleasure vs pain	preference satisfaction vs preference frustration	objective goods
≠ autonomy	≠ autonomy	∈ autonomy?
Intuition: How good/bad is someone feeling?	**Intuition:** How well are someone's wishes, desires etc satisfied?	**Intuition:** How much objective goods from the list does someone have and to what degree?

autonomy is not intrinsically valuable

autonomy might be intrinsically valuable, depending on the list

IS THAT A PROBLEM?

Is autonomy extrinsically valuable?

hedonism:

is autonomy valuable in virtue of bringing about pleasure or diminishing pain?

in general: prima facie yes, at least in some circumstances, namely when autonomy helps you to raise your level of wellbeing.

in the case of nudging technologies: if you have a technology that makes you adopt behaviours that lower your overall wellbeing, then this is bad for you

preference theory:

is autonomy valuable in virtue of bringing about preference satisfaction or diminishing preference frustration?

in general: prima facie yes, at least in some circumstances: people prima facie often have the desire to be autonomous

in the case of nudging technologies: if you have a technology that makes you adopt behaviours that lower your overall wellbeing, then this is bad for you, because usually you prefer not to be harmed

IS THAT A PROBLEM?

Is it ok to take someone's autonomy away with nudging?

rationally ok?
(i.e. in my best interest)

morally permitted?

as a developer?	as a society?	as a user?	depends on moral theory
if the developer is concerned about the wellbeing of their users: depends else: everything goes! $$$	depends: prima facie only if the technology makes people in general adopt behaviours that are overall beneficial to society (yes to fitness trackers, but no to the infinite scroll)	depends: only if the technology makes you adopt behaviours that are overall in your interest (maybe fitness trackers, but no to the infinite scroll)	certainly not morally permitted if the technology is outrightly harmful (no to infinite scroll)