
You’re inside a pleasurable room that you don’t want to leave.
Silently, someone locks the door from the outside.
Are you free?

Thought experiment: *What is human freedom?*



Are you free?

No, according to John Locke (*Essay Concerning Human Understanding*, 1689), because you cannot change preferences and purposes (without encountering confinement).

Genuine human freedom → ***Ability to change what you want, and who you are***

Autonomy (personal freedom) thought experiment

Your psychological profile analyzed by hyper-accurate LinkedIn to assign perfect job opportunities - better than you could find on your own.

But, entering the system reveals your data to all employers, and they will not offer positions outside your digital profile.

So, you get the right job, but you're chained to it, you can't try anything else.

Do you sign up? (Does this increase your freedom?)



Autonomy (personal freedom)

Different *kinds* of freedom?

Get what you want versus *change* what you want

Operational versus structural

Autonomy (personal freedom)

What is the *value* of autonomy, of personal freedom?

How can it be described?

How can it be measured (What are the units)?



a seemingly banal but actually profound and strange experience that has become common in this era: You go to an Airbnb and turn on the TV, already open to someone else's account, and you see all this *stuff* you didn't even know existed. Same TV, same app, same subscription, the house might even feel similar your own — but the screen gives way to an alien world.

Experience of autonomy

(Structural versus operational)



Experience of autonomy

(Structural versus operational)

Outline

What is autonomy? Experience

What is autonomy? Definition

Is autonomy always important?

2 real cases of autonomy dilemmas

Algorithm design to maximize autonomy: Explore/Exploit

Challenges to autonomy

Definition autonomy (Freedom, Self-determination, Agency)

Definitions

- Philosophical
- Social
- Practical



ChatGPT: "Please depict the concept of autonomy in AI ethics with an image."

Definition autonomy (Freedom, Self-determination, Agency)

Philosophical (Aristotle)

- Autos ‘self’ + nomos ‘law’ = Ruling oneself
  - Between extremes of ruled and no rules
  - Between slavery and randomness
  - Between others’ purpose and no purpose

Self-determination

Definition autonomy (Freedom, Self-determination, Agency)

Social (Locke)

Freedom maximization / Freedom limited only by freedom:
Freedom to do what you want… up to the point where it interferes with others doing what they want





Definition autonomy (Freedom, Self-determination, Agency)

Practical

Physically free:
Not constrained/compelled

Informationally free:
Not misinformed, (not uninformed?)

Intentionally free:
Not drugged, altered
Psychologically reasonable



Outline

What is autonomy? Experience

What is autonomy? Definition
    Self-determination / own rules

Is autonomy always important?

2 real cases of autonomy dilemmas

Algorithm design to maximize autonomy: Explore/Exploit

Challenges to autonomy

Autonomy thought experiment #2

(Is autonomy always important?)

Poirot, final season,
final episode,
final moment



What is the autonomy dilemma?

Autonomy thought experiment #2

If you want to advocate for ethical autonomy, do you allow this? Prohibit it? Something else?

(Or, is this a trivialization of principles…)



Autonomy

One choice, but the *right* choice, better than you could find on your own.

Why yes, or why no, in terms of personal freedom?



RISK LEVELS

Movies	N/Y	86%	14%
Jobs	N/Y	90%	10%
Partner	N/Y	79%	21%

FUNDAMENTAL VALUES

NO: Autonomy/Freedom
(Not want to be manipulated, even for own good)

YES: Happiness/Pleasure/Convenience
(If it feels good...)

Autonomy thought experiments #2/3 (Background, ethics risk levels)

The EU AI Act divides AI systems into four ethics categories -
levels 2 & 3 strictly regulated and require AI ethics audit clearance:

- Unacceptable-risk AI systems (Social scoring live with FR)
- High-risk (Medical devices, cars)
- Limited-risk (Chatbots)
- Minimal-risk AI systems (Recommenders, Netflix)

How do you divide acceptable/unacceptable autonomy risks?

Is autonomy like sand or like air?

Outline

What is autonomy? Experience

What is autonomy? Definition
    Self-determination / own rules

Is autonomy always important?

2 real cases of autonomy dilemmas

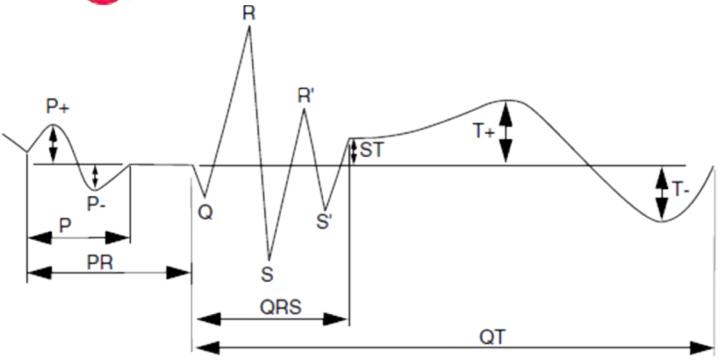
Algorithm design to maximize autonomy: Explore/Exploit

Challenges to autonomy

Autonomy dilemma: Cardisio case

Perfect for AI...

CARDISIO
Your heart counts



Autonomy dilemma: Cardisio case

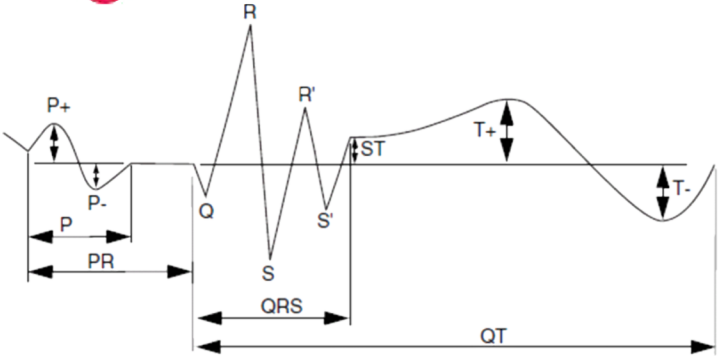
Possible Questions:

Medical:
Does the tool detect heart disease early and save lives?

Financial:
Does the tool save healthcare money?

Ethical:
Does Cardisio increase user autonomy?

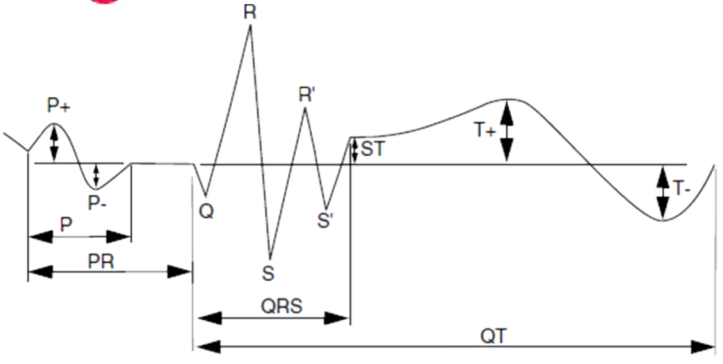
CARDISIO
Your heart counts



Autonomy dilemma: Cardisio case

CARDISIO
Your heart counts

We asked: Does Cardisio increase user autonomy?
(*versus* biological health and high sensitivity relative to specificity)



Autonomy dilemma: Cardisio case

Dilemma #1:

Original output = Green/Red...

For autonomy maximization:

What should outputs be (percentages, colors...)?

Add yellow?

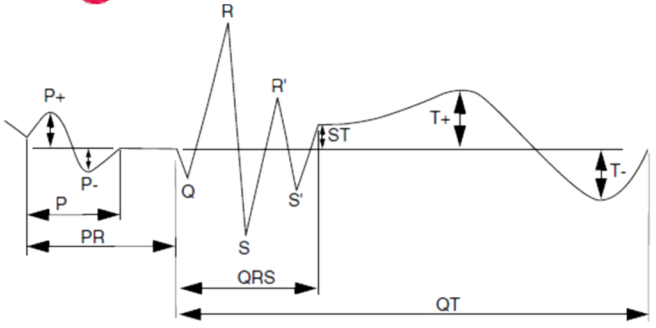
Where are the thresholds?

For whom?

*If you just return the %s, what do those numbers mean? When should you change your life?

CARDISIO

Your heart counts



Autonomy dilemma: Cardisio case

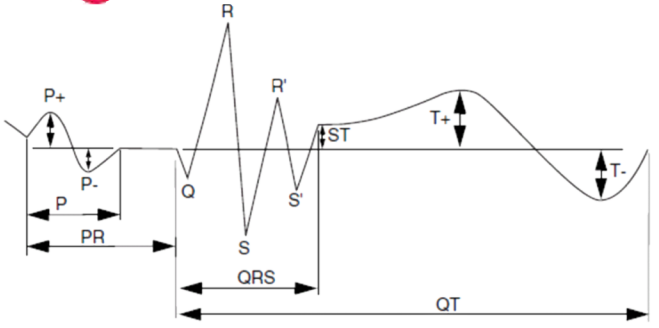
Dilemma #2:

You cannot unknow...

Do you *want* to know?
(If 99% certainty of 99% probability of death in 9 days, do you want to know?)

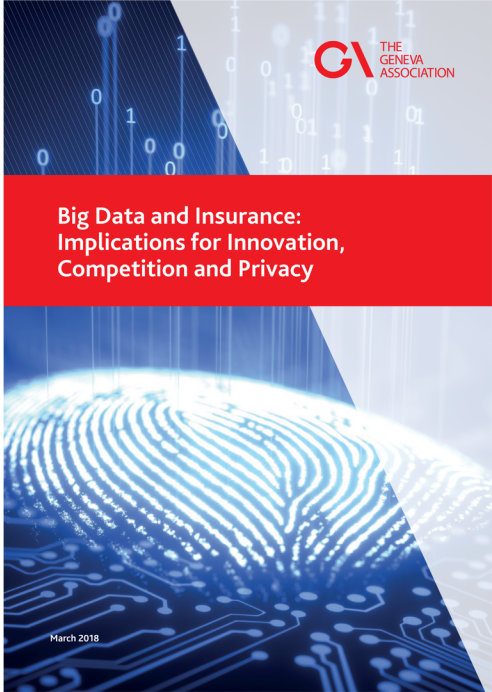
Will you lie to users?

CARDISIO
Your heart counts

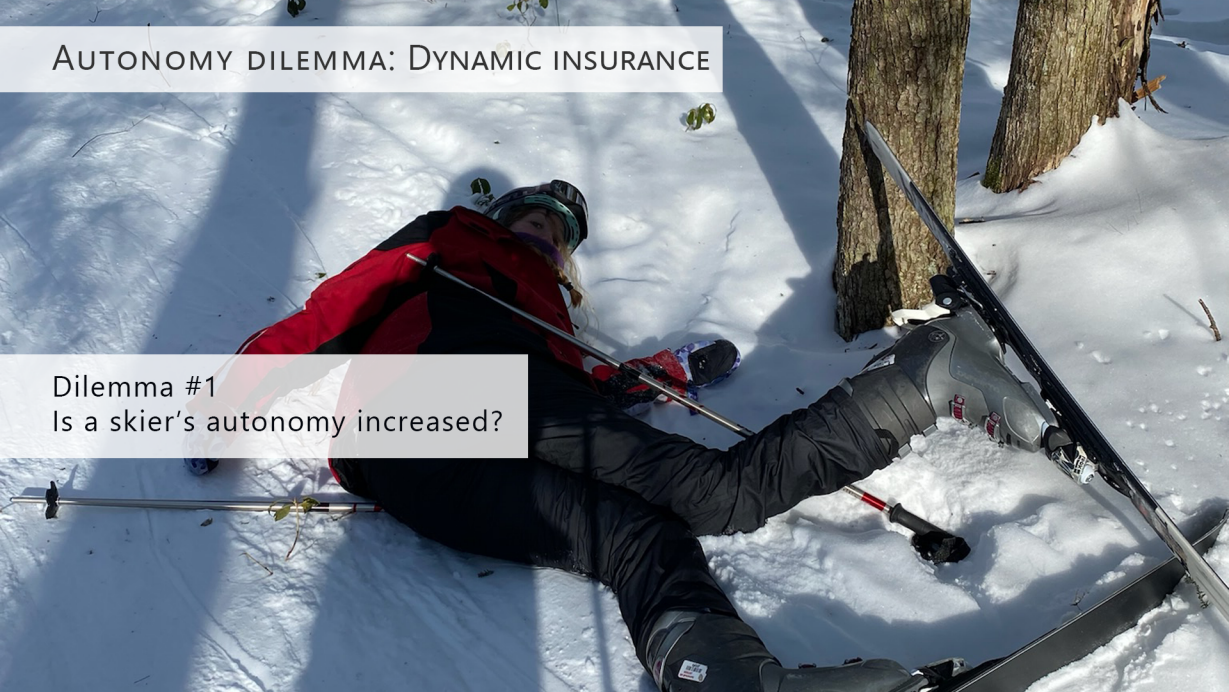


Autonomy dilemma: Dynamic insurance

The ubiquitous availability of vast amounts of data and the ability to analyse it allow for individual and dynamic risk assessment and a continuous feedback loop to policyholders, with no or limited human interaction. By providing risk insights to policyholders, such 'digital monitoring' encourages behavioural change to reduce risks. Moreover, new data sources allow for the implementation of advanced risk management systems that use predictive analytics as a basis for early intervention and risk prevention. Ultimately, these technologies allow the role of insurance to evolve from pure risk protection towards predicting and preventing risks.

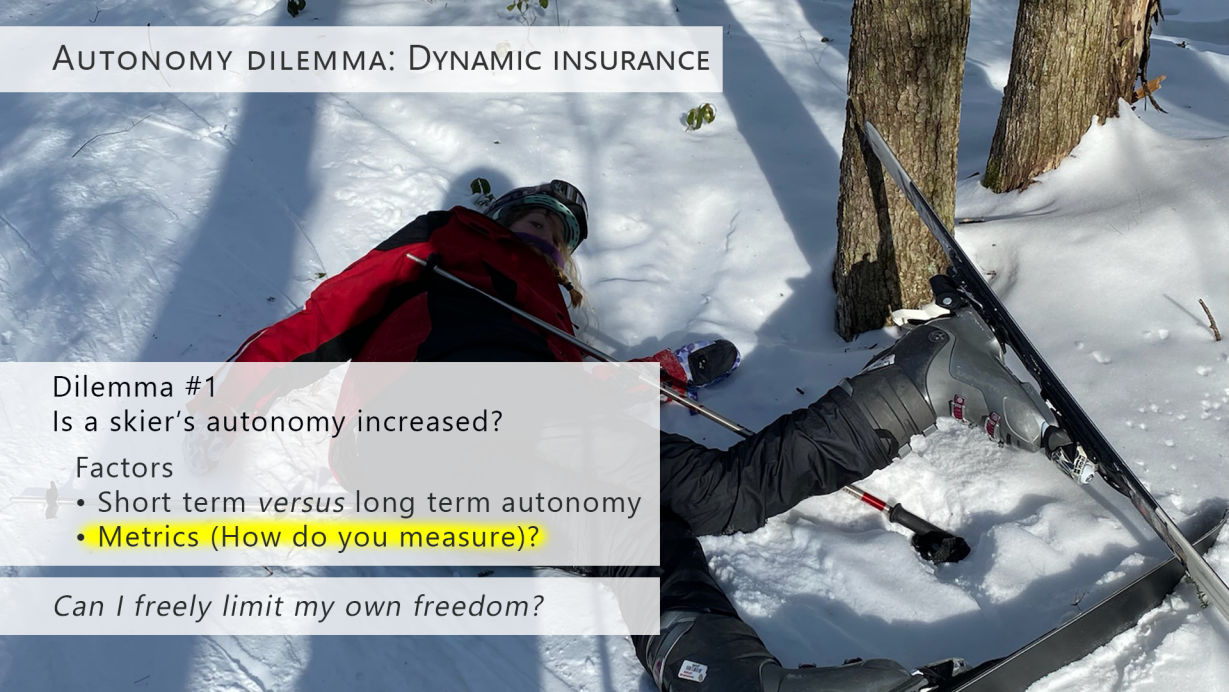


Autonomy dilemma: Dynamic insurance



Dilemma #1
Is a skier's autonomy increased?

Autonomy dilemma: Dynamic insurance

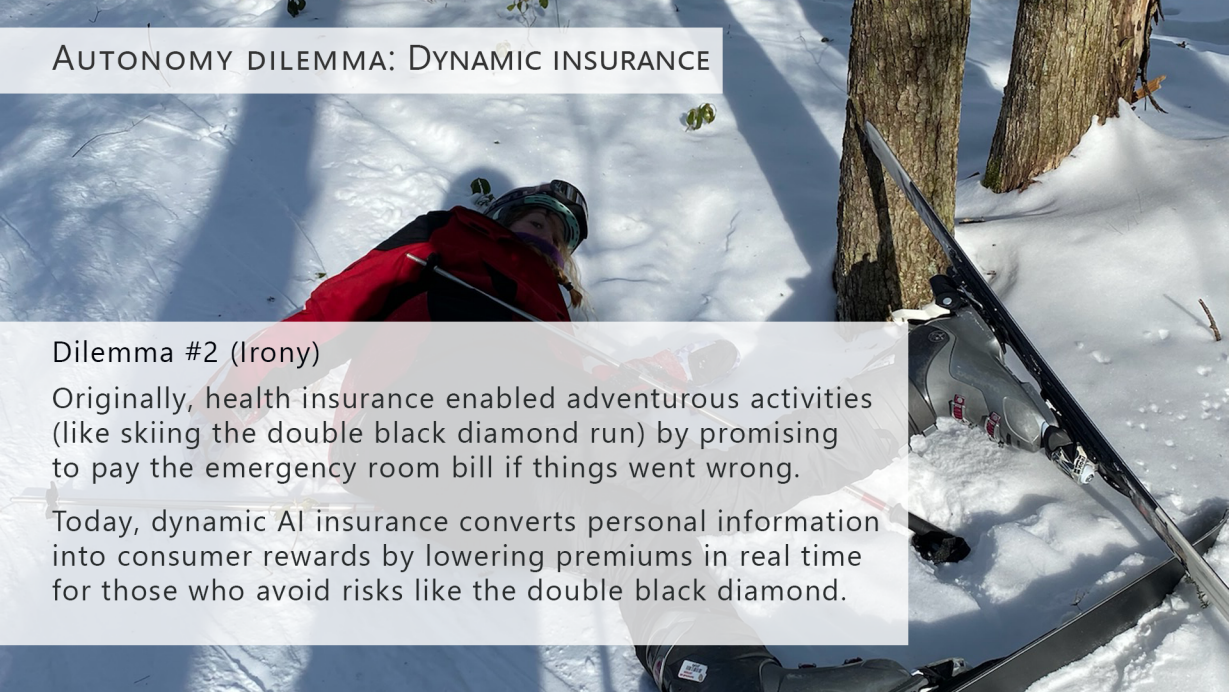


Dilemma #1
Is a skier's autonomy increased?

Factors
- Short term *versus* long term autonomy
- Metrics (How do you measure)?

Can I freely limit my own freedom?

Autonomy dilemma: Dynamic insurance



Dilemma #2 (Irony)

Originally, health insurance enabled adventurous activities (like skiing the double black diamond run) by promising to pay the emergency room bill if things went wrong.

Today, dynamic AI insurance converts personal information into consumer rewards by lowering premiums in real time for those who avoid risks like the double black diamond.

Outline

What is autonomy? Experience

What is autonomy? Definition
    Self-determination / own rules

Is autonomy always important?

2 real cases of autonomy dilemmas

Algorithm design to maximize autonomy: Explore/Exploit

Challenges to autonomy

Algorithm design to maximize autonomy: the Netflix challenge

Engineers can optimize for profit, for time on platform...

Ethics challenge:
How can (Netflix) engineers optimize for user autonomy?



"Illustrate the idea of serendipity as it relates to the search for a movie on Netflix. Use no letters or words. Please illustrate artistically and conceptually, not literally." (ChatGPT, 2024)

Algorithm design to maximize autonomy: the Netflix challenge

Ethics challenge:
How can (Netflix) engineers optimize for user autonomy?

Behavioral challenge:
Give viewers what they want, without trapping by giving the same thing endlessly

Provoke curiosity, new interests, even new personal identities...



Algorithm design to maximize autonomy: Background/Frame



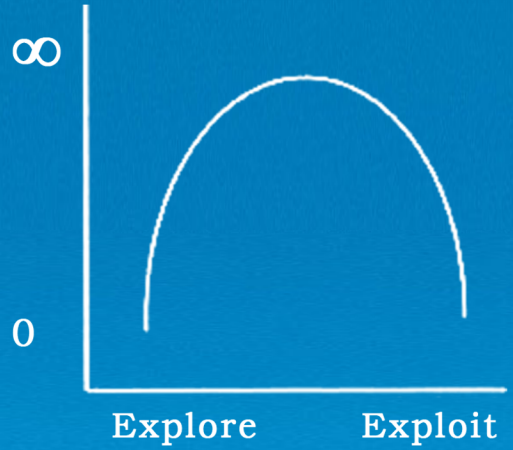
Explore/Exploit - One-armed bandits

"Two people playing one-armed bandit slot machines in a las vegas casino. They are next to each other. Their faces are visible. One person is envious. One person is joyful because he wins a big pot."
(Many regenerations, early 2022, Dall-E)

Algorithm design to maximize autonomy

Netflix behavioral challenge, algorithm optimized for:

- Accumulating platform profit
- Lengthening user engagement

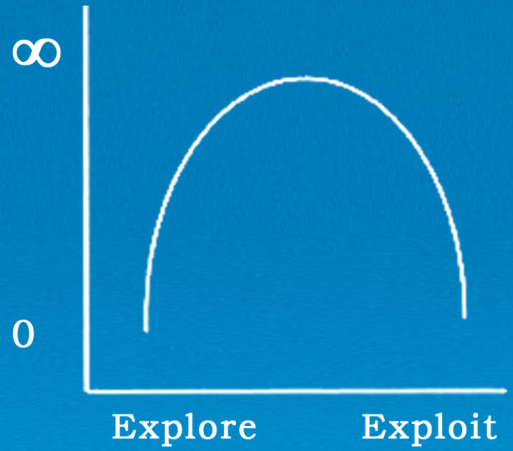


Algorithm design to maximize autonomy

Netflix ethics challenge,
algorithm optimized for autonomy/
human freedom: *What does this mean?*

Where do you tell the recommendation engine to look?

How?



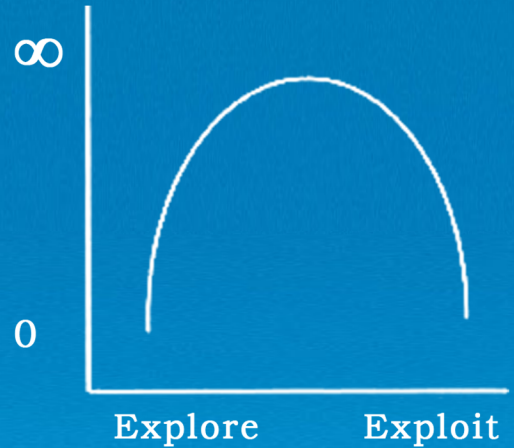
Algorithm design to maximize autonomy

Netflix ethics challenge,
algorithm optimized for autonomy

What would that mean in terms of ***what the user sees when scrolling?***

> Possible direction: literature on breaking filter-bubbles in social media...
> Serendipity studies/nomadism

Are you trying to find something that is similar but different, or different but similar?



Outline

What is autonomy? Experience

What is autonomy? Definition
    Self-determination / own rules

Is autonomy always important?

2 real cases of autonomy dilemmas

Algorithm design to maximize autonomy: Explore/Exploit

Challenges to autonomy

Challenges to autonomy

Nudges (against autonomy?)
  Justifiable nudges?

Dark Patterns
  Dark patterns and psychological/physical control

Nudge (against autonomy?)

"Any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives (ie offering coupons)."
Thaler & Sunstein *Nudge* 2008

+

Behavioral economics

=

Exploit intuitions, biases, emotions, social standing, choice environment: supercharged by AI.



Case: Nudge (against autonomy?)

Examples of how robots could behave include **verbal prompts to donate money**. If a robot possesses the capacity to construct a **detailed user profile**, then offering specifically tailored recommendations becomes a distinct possibility. For example, **if the user has a family history of a particular illness (e.g., heart disease), the robot could suggest contributing to an associated charity (the American Heart Association)**.

Borenstein, J., Arkin, R. Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behavior. AI & Soc 32, 499–507 (2017). https://doi.org/10.1007/s00146-016-0684-1

Is this an autonomy violation?
(If so, is it justifiable? Does it matter if it is presumably a good cause)

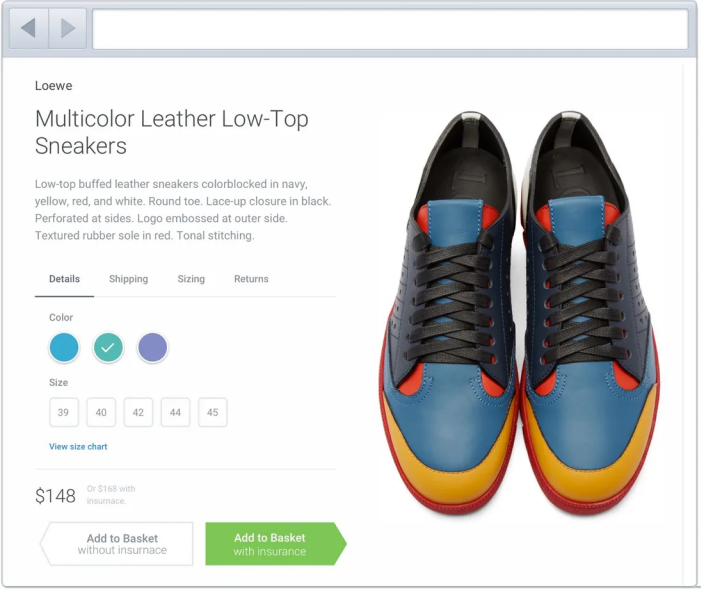
Dark patterns *against* autonomy

User interface design choices that benefit an online service by coercing, steering, or deceiving users.

- Arvind Narayanan *et al.*, *Dark patterns at Scale*

Misleading links, boxes pre-checked for spam, phone numbers requested for one purpose and used for others.

(A/B tested, possibly supplemented by personal data)



Nudges/Dark patterns at the limits

Physical or Psychological? (AI as drug dispensation)

"The thought process that went into building these applications, Facebook being the first of them, was all about: 'How do we consume as much of your time and conscious attention as possible?'

That means that ***we need to sort of give you a little dopamine hit every once in a while***, because someone liked or commented on a photo or a post or whatever.

And that's going to get you to contribute more content, and that's going to get you... more likes and comments."

-Sean Parker, Facebook's founding president,



Outline

What is autonomy? Experience

What is autonomy? Definition
    Self-determination / own rules

Is autonomy always important?

2 real cases of autonomy dilemmas

Algorithm design to maximize autonomy: Explore/Exploit

Challenges to autonomy

Postscript: Control *versus* disciplinary (spaces) society

