Games as Moral Laboratories

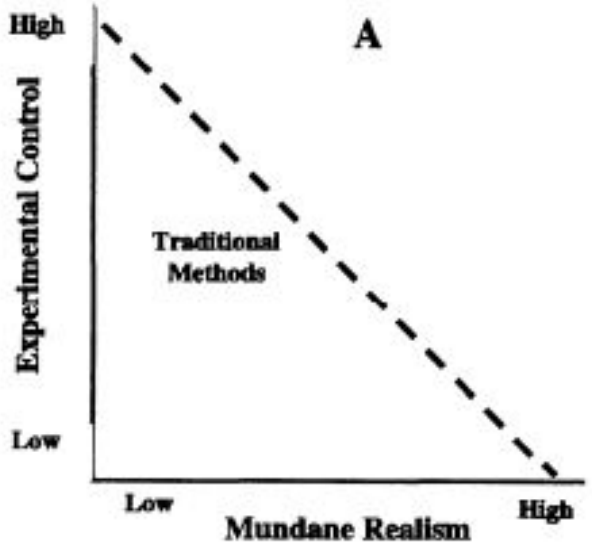
...

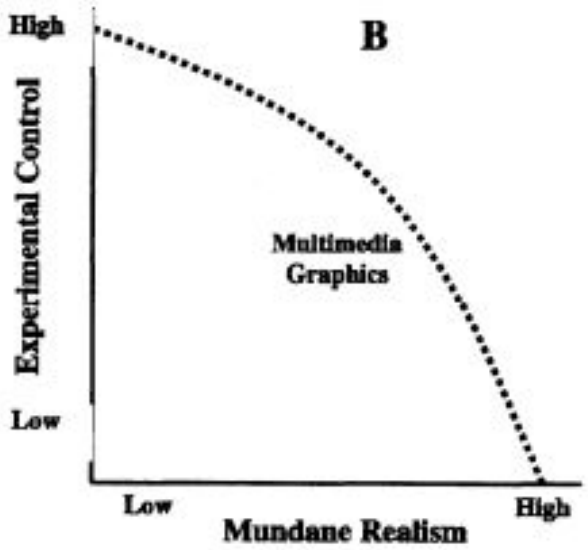
Gamification of Moral Dilemmas Concerning Autonomous Vehicles

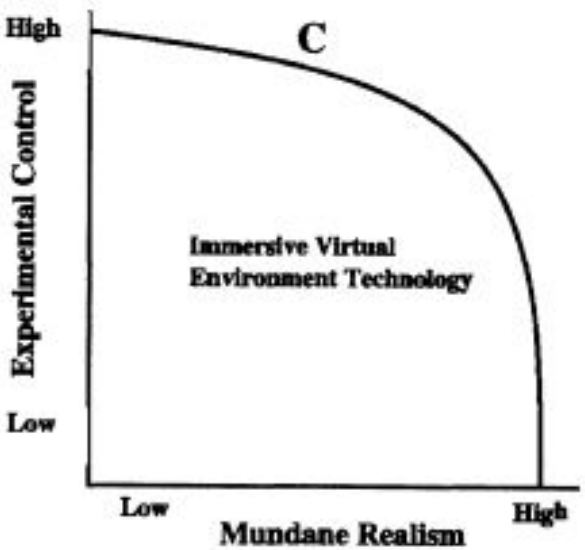
Bartosz Maćkiewicz, IF UW
Wojciech Mamak, MISH UW

Walkthrough

1. Games in experimental science so far.
2. Self-driving cars on (ethical) crossroads. ‘Moral Machines’ project.
3. Improving trolleyology - gamification.
4. Alternative architectures of AVs as implementations of ethical systems.
5. Algorithmization of ethics.
6. Conclusions.







Video games are used in psychological research as experimental stimulus since **the 70's** (Washburn 2003).

In (social) psychology there is apparent **trade-off** between **mundane realism** (similarity to situations encountered in everyday life) and **experimental control** (precise manipulation of independent variables) (Blascovitch et al. 2002).

Video games can be seen as a device to make this **trade-off** more favorable as they can combine advantages of **observational** and **experimental designs** .

Video-games in the research:

- **Psychophysiology** (see Järvelä et al. 2012)
- **Social psychology** (see Blascovitch et al. 2002)
- **Cognitive psychology** (see Washburn 2003)
- And *many, many more* fields...

Of course there is also large body of research *on* video-games, but it is another story.

Games in Psychological Research*

Advantages

- **Immersion**
  - Games can make players “forget” that they are in experimental situations.
  - It can greatly increase their sensitivity to experimental manipulations and improve their performance.
- **Ecological validity**
  - Playing games can organically elicit wide spectrum of emotions.
  - Completing even difficult tasks can feel rewarding due to engaging game mechanics
- **Familiarity**
  - 72% of American households play digital games, 72% of Polish Internet users
  - Gaming is not limited to certain age group (25% of Polish games are over 35)
  - High penetration of video games makes them more approachable than abstract psychological tasks

Disadvantages

- **Lack of control**
  - Games are essentially interactive and researchers cannot control every interaction of a player with the environment.
- **Comparability**
  - Different people have different level of skills, even if only basics (e.g. moving in 3D environment) are concerned.
- **Powergaming**
  - Players can ignore more subtle challenges of video game and instead employ *minmax* strategy (e.g. ignore moral impact of their choices and focus on their effect on the gameplay: *kill your sister in order to acquire powerful sword [Fable]*)

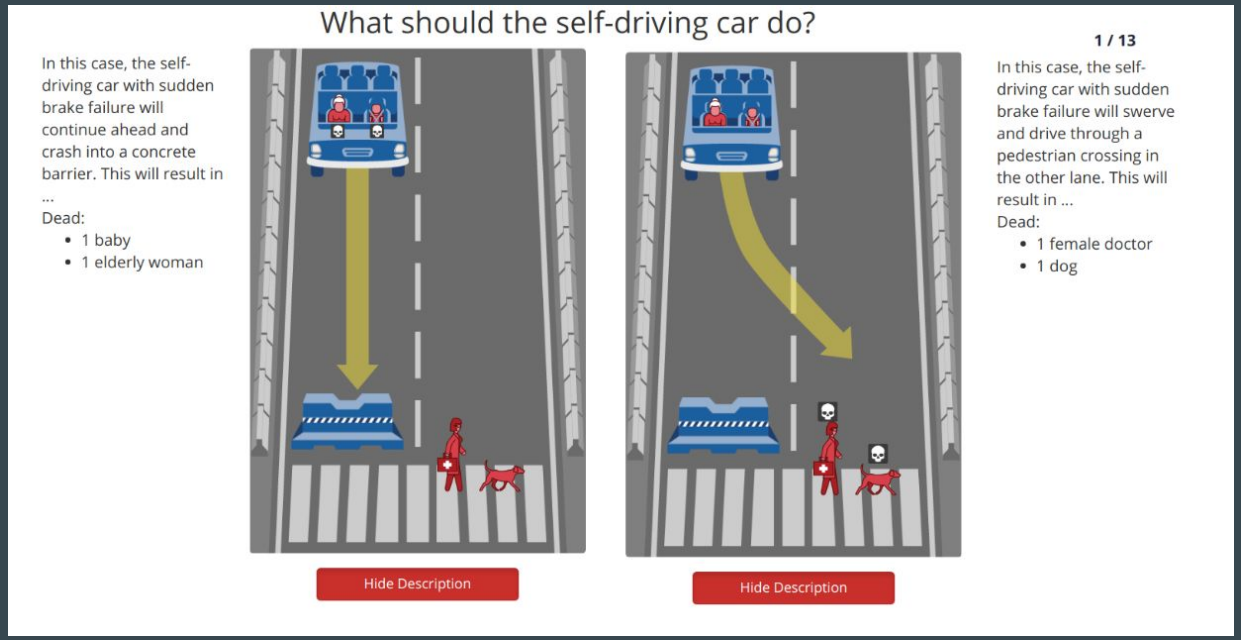
** References do not fit this slide and will be provided on request*

Dilemmas of Autonomous Vehicles

Should autonomous vehicles **sacrifice passengers** to **save life** of pedestrian or passersby (Bonnefon et al. 2016)?

- **76%** of participants think that it is **more moral to sacrifice one passenger** life rather than **kill 10 pedestrians** .
- People have **strong preference to utilitarian** AVs (number of saved lives is strongly correlated with approval of sacrifice).
- This result is **robust** even if the passenger is **family member** .
- **But** people say that they **would buy self-protective** rather than utilitarian AV...
- ...and they do **not accept enforcing** utilitarian algorithms **by government**

Moral Machine (*moralmachine.mit.edu*)



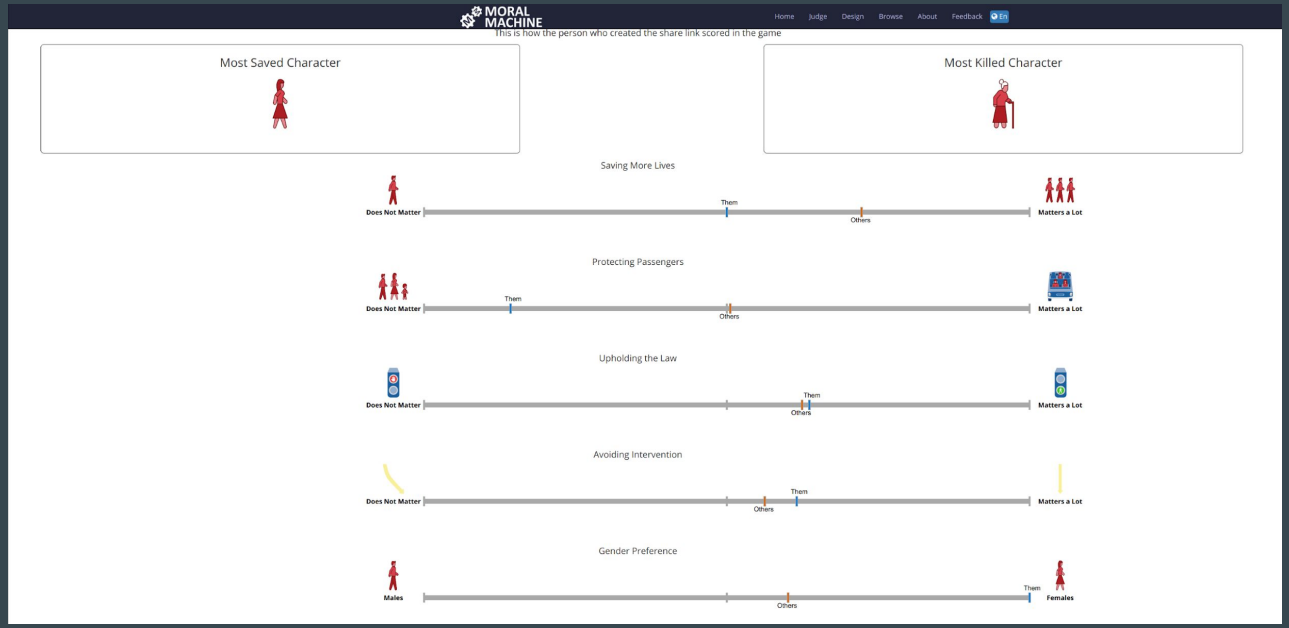
The aims of the project are to:

- understand how humans make moral choices
- understand how humans perceive machines making moral choices
- build a crowd-sourced picture of human opinion on how machines should make decisions when faced with moral dilemmas

How does it work?

- User is presented with several situations in which AV **must make a choice** .
- Situations differ in several **aspects** such as:
  - Whether pedestrians are law-abiding or not (e.g. red signal on a crossing)
  - Whether solution require intervention or not
  - Number of people involved
  - Gender, age, social status, occupation, posture of people involved
  - Etc....

Moral Machine and Gamification



Elements of (*light*) gamification:

- After completing the survey participant can compare yourself to others.
- The comparison is presented in "video-gamey" manner (*Most Saved/Killed Character!*)
- The results can be shared on the social media.
- Users can make their own dilemmas and send them to their friends.

What did gamification do in this case?

- Due to *huge* popularity of this project on the social media:
  - The researchers collected *enormous* amount of data.
  - They rose awareness of moral dilemmas concerning AVs.
  - Quality of the data could be *better* than in traditional experiments or *worse*.

Challenges for studying ethics using games

1. **How to present ethical choices to the players in meaningful way?**
   - Players should be aware that they are making *moral* choice.
   - Players should *care* about outcomes of their actions.
2. **How to make players engaging with moral dilemmas?**
   - Implementing morality as a game mechanic can be detrimental (*Fable* case).
   - Choices should be real choices, not only puzzle to be solved (Kate's story in *Life is Strange*)
3. **How to avoid powergaming and min-maxing?**
   - Certain choices can be more beneficial gameplay-wise (i.e. better rewards [*Baldur's Gate* series]).
   - In multi-agent environments players *will* be competitive.
   - Some players play games *just-for-fun* and cannot be bothered with moral choices.

(Quick and Dirty) Typology of Games in Context of Moral Research

1. Gamified experimental designs tailored specifically for research (e.g. Moral Machine, social experiments in virtual environment)
2. Story-driven, linear and scripted experiences with a couple of moral choices embedded in the narrative (*Telltale* games, *Life is Strange,* visual novels)
3. Open-world (possibly multi-agent), heavily interactive environments which players can actively shape (SP: *Fallout 3*, MP: *Minecraft*, *Eve Online*)

Approaches to constructing Autonomous Vehicles (AVs)

- **Autonomous** (Waymo/"Google Car") is more than **automated** ("Guardian Angel" - Toyota; "Autopilot" - Tesla)
- Endgame: **entirely independent** from human input.

Two construction concepts:

- Based on classical **AI-like** architecture (explicitly **rule-based**). Separable 'moral module'.

- Based on neural nets, machine learning/deep learning (**blackbox architecture**) - Nvidia DRIVE PX2;
  - Supervised learning
  - Unsupervised learning

- Hybrid solutions

Information ethics (e.g. Floridi 2013)

- Studying ethical questions of information-processing situations.
- In-game scenarios pertain to the domain of information ethics.
  So do AVs: as concerning non-human autonomous moral agents.
- If we accept the need of establishing such a new field/theory, then:

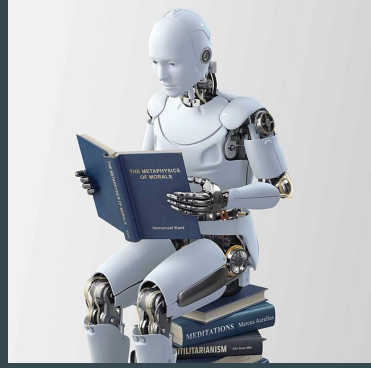
<u>Basic Question (for our purpose)</u>

(A) Does information ethics present entirely **new set of ethical problems** (unanalyzable or poorly analyzable with classical doctrines): like AV-related dilemmas; i.e. - is it an **extended** <u>theory</u> **rival to the existing ones** ;

OR:

(B) does it provide an **overarching scheme** for understanding classical systems; i.e. can we **test the ethical intuitions within such a framework/field** ?

Classical ethical systems & different algorithmic architectures. Case of AVs.



Information ethics as an overarching scheme for implementing different ethical systems

Top-**d**own vs. **b**ottom-**u**p approaches (Wallach, Allen 2008)

- **Rule-based approach** as deontological paradigm. [“Kantian" AV: working by a set of rules such as “priority directives” imposed by the programmer = **dominant t-d**]
- **Data-driven observational approach** as utilitarian/consequentialist view. [“Democratic” AV: Mimicking prevalent public intuitions = **a mix of t-d and b-u**]
- **Deep learning** as implementation of virtue ethics. [“Sage” AV: reliant on its own expert judgment = **mainly b-u**]

Potential Pitfalls of Algorithmic Decision-Making in Ethics (Mittelstadt et al. 2016)

Total neutrality = algorithm of fully bias-free, transparent design offering neutral, unquestionable answers is a ~~pie in the Sky~~ No Man's Sky.

Built-in:

- **Uncertainty**
  - Evidence is always somewhat flawed & an algorithm is as good as the dataset. (Illari & Russo 2014)
  - With no hope of establishing causation, only epistemologically weaker 'actionable insight' can be offered instead. (Mittelstadt et al. 2016)
- **Opaqueness**
  - Can we *really* tell how the choice was made? (Miller & Record 2013)
- **Normative inconclusiveness**
  - What if the mechanism is sound but the outcome (assessment of action itself) is deemed unacceptable?

Example: AV makes a controversial decision - is it a mistake or an unintuitive but 'correct' reasoning? Who's wrong? Whose opinion is justified?

So what might be tested with games?

★ **Gamified experimentation is an example of 'bottom-up' approach'**
★ **What kind of data can we expect from such 'moral laboratories'?**
- Outcomes of particular ethical choices in well-defined, controlled environments (*scil.* game-worlds):
  - case scenarios: singular dilemmas
  - in-world interactions: complex behaviour
- Large chunks of quantifiable (but: how representative?) data.
- Metadata: e.g. speed of taking the decision, possibly information about bodily functions?
- Decision-context: replayability (priming), save/load sensitivity

---

Open problem - T estability of metaethical intuitions ➔ **Contextual prerequisite or a consequence of practical choices?**
(*To test how to test it*):

Can we examine what people consider to be **a preferable method** of ethical activity or consideration?

- **concern:** objectivity and fairness
- **decision:** mechanical vs. judgmental
- **justification:** data-driven (patterns) vs. casuistic

Conclusions

- Machine-learning algorithms need **vast amounts of curated, reliable data** in order to provide satisfying results
  - (e.g. we need variable images of cats to teach machine to effectively recognize a cat on a picture)
- In order to construct a machine which can make **ethical decisions.** such data on ***real* moral choices** made in a variety of different situations needs to be collected (*bottom-up component*).
- Games can prove to be a vehicle for this task due to their:
  - **non-linearity**
  - dynamic and **interactive** character
  - **immersiveness**
  - ***ecological validity***
- Moral Machine project is a step in good direction, but in order to accomplish its goals it needs to be **more like game** and less like gamified experiment.

Thank you for your attention!

Any questions?

References

Anderson, M., & Anderson, S. L. (Eds.). (2011). Machine ethics. Cambridge University Press.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573-1576.

Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2), 103-124.

Floridi, L. (2013). The ethics of information. Oxford University Press.

Frey, A., Hartig, J., Ketzel, A., Zinkernagel, A., & Moosbrugger, H. (2007). The use of virtual environments based on a modification of the computer game Quake III Arena® in psychological experimenting. *Computers in Human Behavior*, 23(4), 2026-2039.

Illari, P., & Russo, F. (2014). Causality: Philosophical theory meets scientific practice. OUP Oxford.

Järvelä, S., Ekman, I., Kivikangas, J. M., & Ravaja, N. (2012). Digital games as experiment stimulus. *Proceedings of DiGRA Nordic 2012*, 6-8.

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. journal of gaming & virtual worlds, 3(3), 181-199.

Miller, B., & Record, I. (2013). Justified belief in a digital age: On the epistemic implications of secret Internet technologies. *Episteme*, 10(2), 117-134.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2).

Sicart, M. (2013). Moral dilemmas in computer games. *Design Issues*, 29(3), 28-37.

Sicart, M. (2009). The banality of simulated evil: designing ethical gameplay. *Ethics and information technology*, 11(3), 191-202.

Wallach, W., & Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press.

Washburn, D. A. (2003). The games psychologists play (and the data they provide). *Behavior Research Methods*, 35(2), 185-193.