Discussion: "Economic analysis of ethical principles for AI algorithms: An autonomous driving case"

Duk Gyoo Kim
SKKU

February 6, 2025

Setup

An autonomous driving vehicle (ADV) version Trolley dilemma: "When ADVs expect an accident, should they be **programmed** in minimizing the total expected harm or protecting passengers?"

- Government sets a rule: r1(utilitarian), r2(Rawlsian), r3(industrial growth or protection of passengers).
- Manufacturer chooses either *a* (protect passengers) or *d* (protect pedestrians). Harms to passengers and pedestrian when unprotected are α and β.
- A citizen is an ADV passenger with probability *p* and a pedestrian otherwise.

Main findings

Taking the average utility of the citizens as social welfare,

- r1 is weakly better than r2 and r3;
- under some economic conditions, r2 and r3 can achieve the same outcome of r1.

What I like about the paper

- Timeliness: Self-driving cars would occupy the streets soon. We need proper discussions.
- A simple comparison of three social rules

What I am less sure about

- Citizens are randomly assigned to be either a passenger or a pedestrian. Don't they choose what they want to be? To be more specific, the government and manufacturer's decisions could affect the citizens' decision of purchasing ADVs.
- Comparing the rules based on the average utility of citizens? The average utility is essentially the (normalized) utilitarian social welfare. It seems natural that r1(utilitarian) is superior in this comparison.
- It is well known that people normatively prefer ADVs to minimize expected harm, while they are more willing to adopt an ADV if it prioritizes the passengers. This discrepancy between normative and individual preferences on ADVs could have interplayed with the government's rule setting.

Minor points

- "rule 1: max social welfare (r1, Utilitarian)" is a somewhat incorrect expression. The utilitarian social welfare is one particular form of measuring social welfare, not a synonym.
- It seems unnecessary to consider a continuum of citizens, unless type heterogeneities are added to the model.
- Can a mixed strategy (probability distribution over a and d) be considered?
- With no enforcement problem, it seems unnecessary to separate the government from the manufacturer.
- Industrial growth (or the protection of passengers) as social welfare needs to be convinced more. It is basically the utilitarian social welfare with zero weights on pedestrians.
- Check Feess and Muehlheusser (2024) "Autonomous Vehicles: Moral dilemmas and adoption incentives"

Even more minor

The shape of utility function would change the socially-optimal outcomes. My favorite exam question goes:

"A narrow bridge connects points A and B, and it takes 10 minutes to cross. When a car is on the bridge, cars traveling in the same direction can enter, while those heading in the opposite direction must wait at the entrance. Consider four self-driving cars: Car1, Car2, and Car3 arrive at point A at 8:00, 8:09, and 8:18, resp., while Car3 arrives at point B at 8:10. In what order should the cars cross the bridge?"

- If their utility function is $u(w) = -w$, where w is a waiting time, the utilitarian social planner lets 1–2–3–4 move.
- If $u(w) = -w^2$, the same utilitarian social planner lets 1–4–2–3 move.