Challenges for Transparency

ICML Workshop on Human Interpretability
Thurs Aug 10, 2017

Adrian Weller
University of Cambridge and Alan Turing Institute

For more information, see
http://mlg.eng.cam.ac.uk/adrian/

Summary of paper

Transparency is often beneficial but it is not a universal good.

- There are many types of transparency with different motivations – we need better ways to measure them.
- We should recognize that sometimes transparency is a means to an end, not a goal in itself.
- Actors with misaligned interests can abuse transparency as a manipulation channel, or inappropriately use information gained.
- In some settings, more transparency can lead to less efficiency, fairness or trust.

What is transparency?

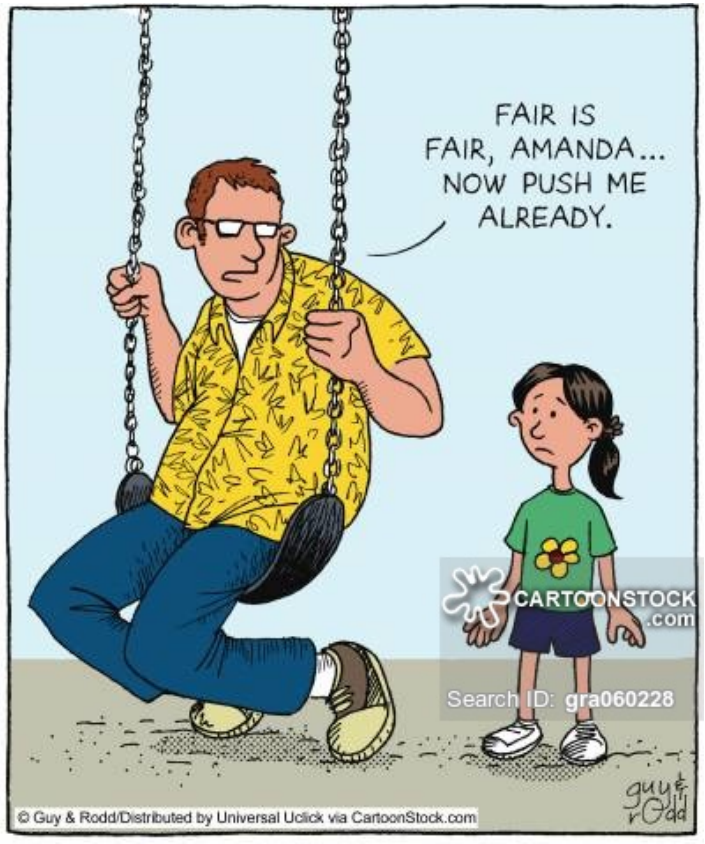
In some ways, transparency is like fairness

- Both are ideas that almost everyone likes –
- Yet both can mean different things to different people in different contexts!

What is transparency?

In some ways, transparency is like fairness

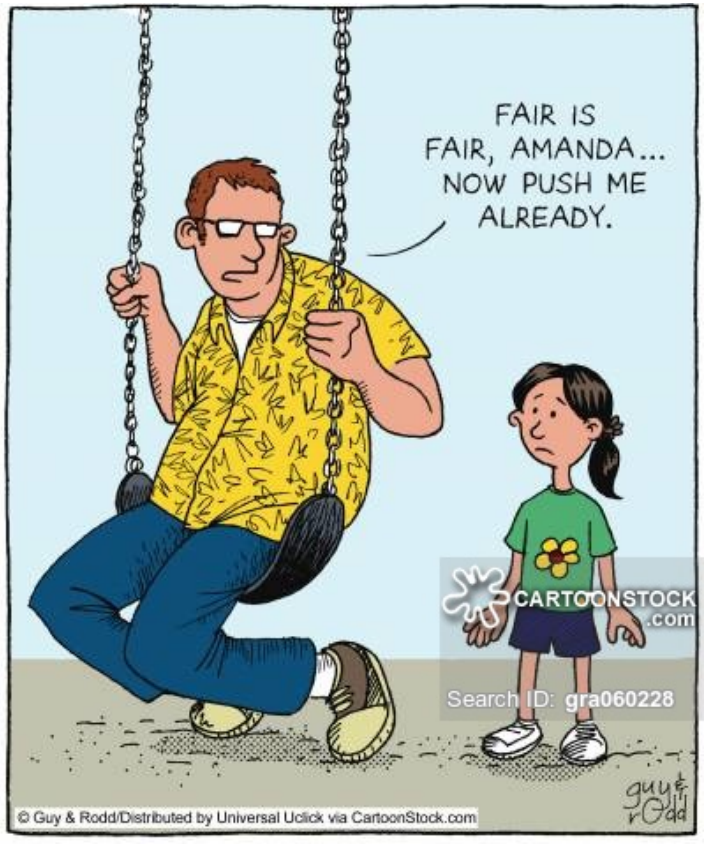
- Both are ideas that almost everyone likes –
- Yet both can mean different things to different people in different contexts!



What is transparency?

In some ways, transparency is like fairness

- Both are ideas that almost everyone likes –
- Yet both can mean different things to different people in different contexts!



What is transparency?

Types of Transparency

1. For a developer, to understand how their system is working, aiming to debug or improve it: to see what is working well or badly, and get a sense for why.

3. For society broadly to understand and become comfortable with the strengths and limitations of the system.

4. For a user to understand why one particular prediction or decision was reached, to allow a check that the system worked appropriately and to enable meaningful challenge (e.g. credit approval or criminal sentencing) – *local interpretability*.

5. To provide an expert (perhaps a regulator) the ability to audit a prediction or decision trail in detail, particularly if something goes wrong (e.g. a crash by an autonomous car) – accountability, liability.

Each type of transparency motivates different measures which can be hard to define precisely.

Types of Transparency: Audience vs Beneficiary

We can differentiate between the intended **audience** of an explanation and the likely **beneficiary** (or beneficiaries).

The deployer owns the system and releases it to the public.

More types of transparency:

7. To make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system. Beneficiary: deployer.

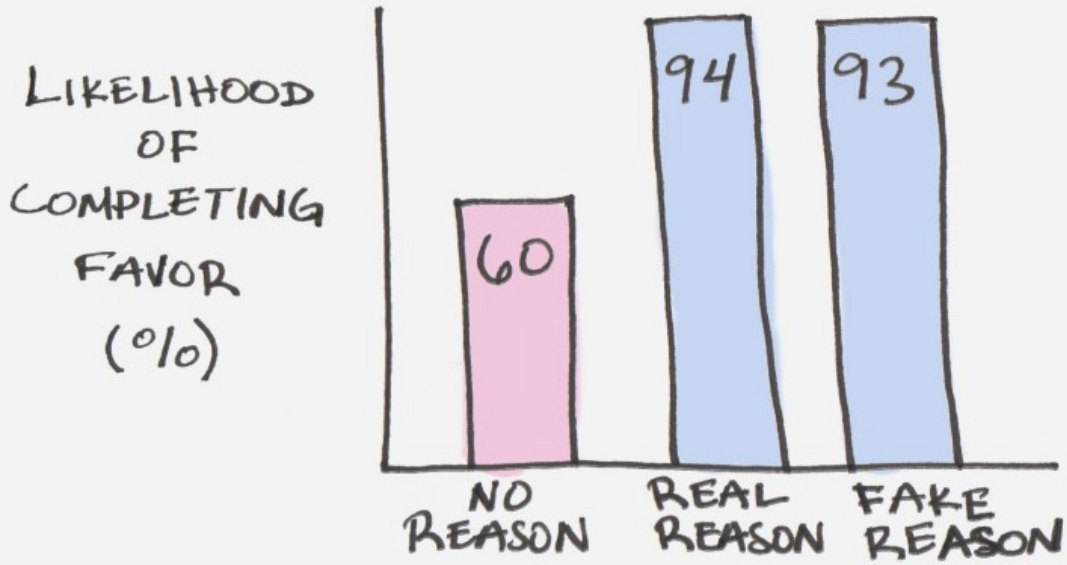
8. To lead a user (the audience) into some action or behavior – e.g. Amazon might recommend a product, providing an explanation in order that you will then click through to make a purchase. Beneficiary: deployer.

Explanation as Manipulation? (Langer et al., 1978)

The Copy Machine Study

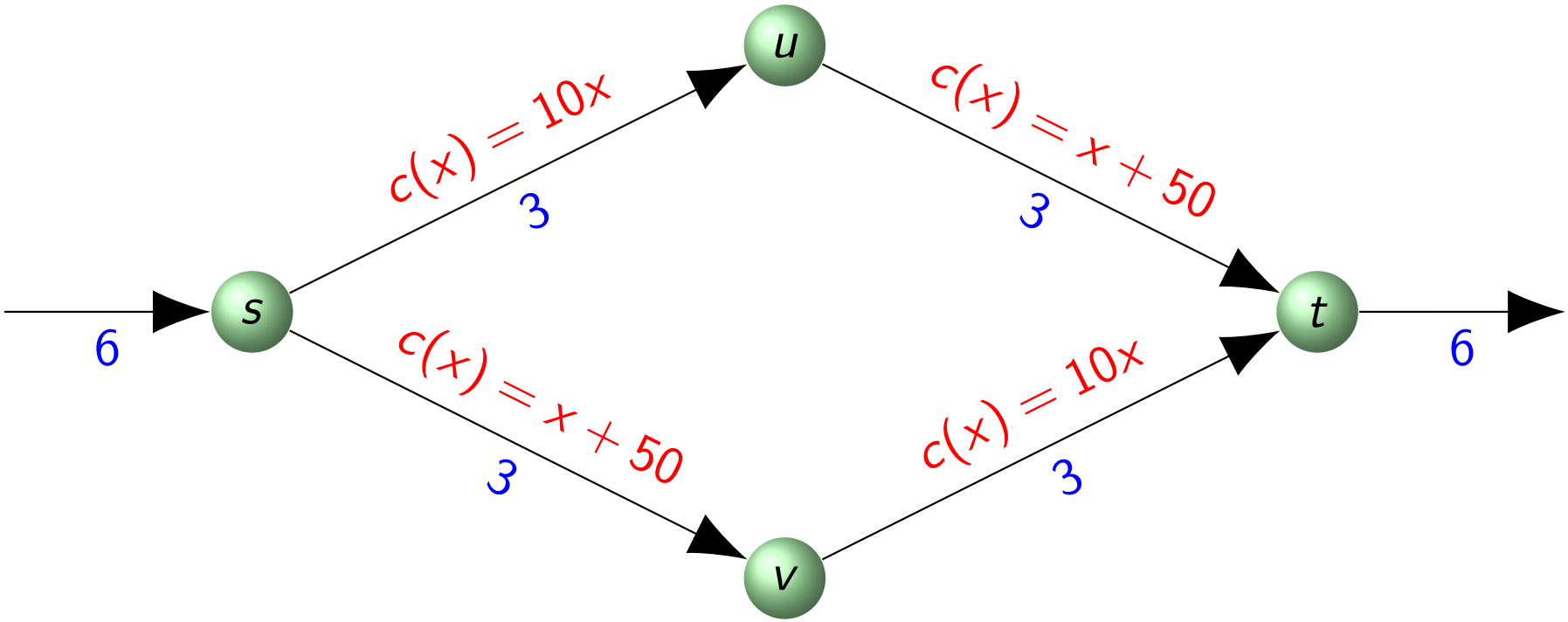
"A well-known principle of human behavior says that when we ask someone to do us a favor we will be more successful if we provide a reason. People simply like to have reasons for what they do."

–Robert Cialdini



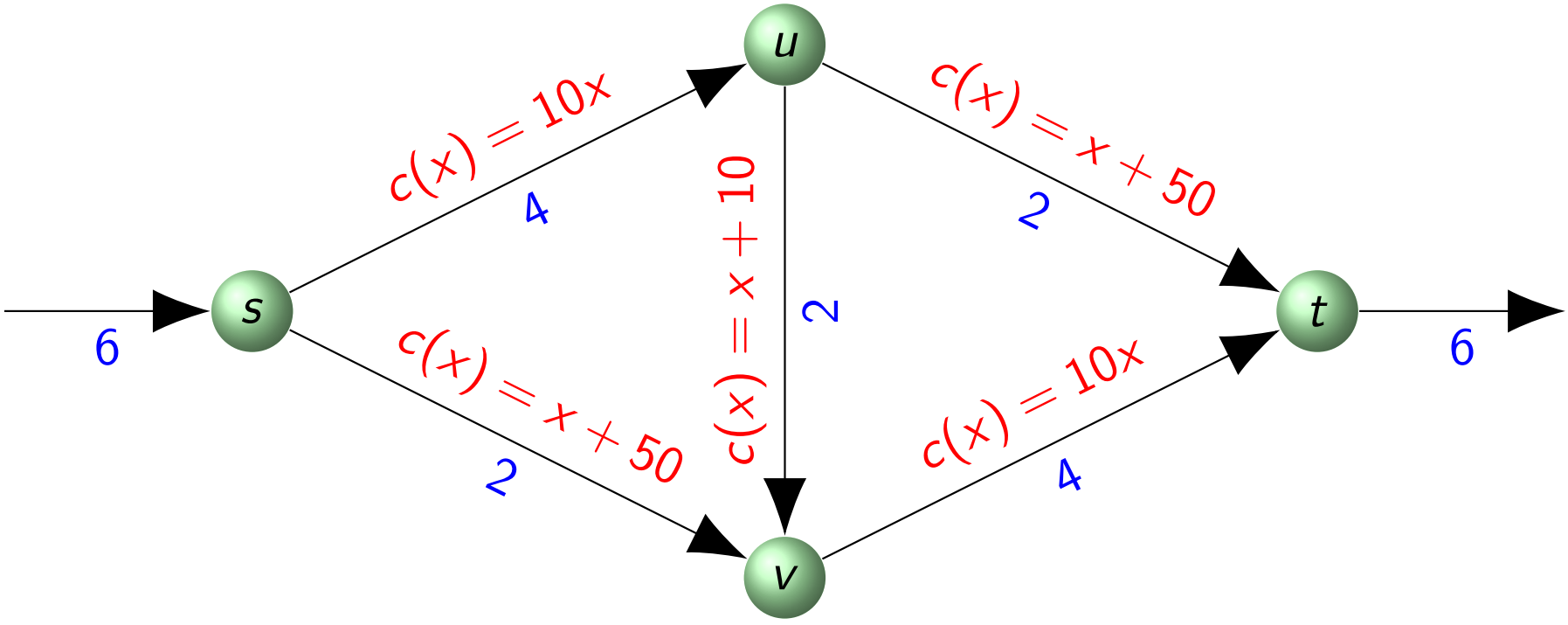
Type 3 transparency: Braess' paradox (Kelly, 1991)

Initial configuration. Everyone has cost (delay) of 83.



Type 3 transparency: Braess' paradox (Kelly, 1991)

Edge $u \to v$ is revealed. Everyone has cost (delay) of 92.



Revealing the extra path – full transparency – makes everyone worse!

Conclusion

- There are many settings where transparency is helpful.
- We begin to clarify just what sorts of transparency may be desirable for each, with accompanying research challenges.
- We highlight scenarios where transparency may cause harm.
- We hope to stimulate research into what sorts of transparency are helpful or harmful to whom in particular contexts.
- This is a rich area: we can draw on connections across economics, multi-agent game theory, law, policy and cognitive science.

Thank you

For more information, see `http://mlg.eng.cam.ac.uk/adrian/`