DIGITAL DYSTOPIA

Game Theory Conference

July 24, 2020

Jean Tirole

Toulouse School of Economics

IAST

INTRODUCTION

Innovation $\Longrightarrow$ low cost of collecting, storing and analyzing personal data.

Will this

- promote a more civilized society?
- lead to a mass surveillance by platforms and governments holding and integrating too much information about what defines us as individuals?

Junction at which the new technology [connected objects, social networks, ratings, artificial intelligence, facial recognition, cheap computer power…] comes to maturity

$\Longrightarrow$ social science fiction is needed to

- understand the channels through which a dystopic society might come about
- design legal and constitutional safeguards.

A MOTIVATION: THE CHINESE SOCIAL CREDIT SYSTEM (2020)

By 2020, individual social score will embody a variety of criteria. ▸ Criteria

Illustrates potential problems. Warning: this social scoring is

- not cast in stone; pilots may differ from future implementation anyway
- probably not a Chinese specificity.

What for?

- Elicit social sanctions by peers: public stigmatization/modern pillory (trend in the US too): C2C. Lead example.
- Enlist corporations and public entities to restrict access to discounts on purchases, employment, transportation, visas abroad, access (of individual or children) to the best schools or universities: B2C.
- Restrict business partners: B2B.

Why does a Leviathan with enough leverage to sustain a law that creates individual social scores not employ more traditional compliance policies such as brute force and imprisonment?

1) For an autocratic state

   - traditional repression is expensive (inefficient and corrupt courts, cost of imprisonment…) when it extends beyond a small minority; Huxley's letter to Orwell concerning *1984* (out to lunch).
   - international opprobrium.

2) Furthermore, the underlying logic may be harnessed not only by autocratic governments, but also by entities with limited coercive power:

   - a majority in a more democratic regime
   - a private platform.

MODELING STRATEGY

(1) Social relationships come in two guises:

- strong ties/stable ones (family, friendship, village or employee bonds)
- weak ties/transient ones (platform/independent contracting or large city interactions)

$\Rightarrow$ An agent's behavior may become known to others in two ways:

- through direct experience of interacting with the agent
- through a publicly disclosed social score aggregating the individual's behaviors.

(2) Agents have image concerns $\Rightarrow$ signal prosocial proclivity (intrinsic motivation to do good/be trustworthy in social interactions)

(3) Commitment to the methodology of construction of social score (information design).

OUTLINE

1. Reminder on social incentives

2. Leveraging social sanctions to promote ruler's political, religious or societal agenda

- Can state design the social score so as to induce more compliance with the ruler's objectives? When will the state do so?
- Can a private platform or a majority in a democracy use similar techniques so as to achieve its goals?

3. Guilt by association Coloring of a person's perception by the company she keeps has become very cheap with face recognition and artificial intelligence

- When will state add yet another social pressure -ostracism- to toe the line?

▸ Related literature

II. THE CALCULUS OF SOCIAL APPROVAL

Modeling ingredients: those in Bénabou-Tirole (2006) and theoretical and empirical literatures on prosocial behavior.

Behavior ($a_i \in \{0,1\}$) is driven by

(1) intrinsic motivation to do good, ve, where
   - $e =$ externality
   - $v \sim F(v)$ on $[0,1]$ (mean $\bar{v}$)

(2) cost of doing good, $c > 0$

(3) desire to project a good image of oneself.

Mass 1 of agents i. Agent i selects an action $a_i \in \{0,1\}$, which affects j (could be multiple actions/multiple partners)

- stable relationships, intensity of image concerns μ (vis-à-vis current partner j)
- transient relationships, intensity of image concerns ν (vis-à-vis future partners).

Information

Silo information: only direct observation

- j observes a_i (could be with noise, here perfectly)
- future partners observe nothing $(\emptyset)$

Transparency: everyone further observes social score

$$s_i = a_i \Rightarrow s_i = \begin{cases} 1 & \text{if } a_i = 1 \\ 0 & \text{if } a_i = 0 \end{cases}$$

Payoff function

$$u_i = (v_i e - c)a_i + \mu \hat{v}_i(I_{ij}) + \nu \hat{v}_i(I_i).$$

silo	a_i	$\emptyset$
transparency	$\{a_i, s_i\}$	s_i

where here $s_i = a_i$.

Welfare

Image = positional good (constant-sum). Total reputation, $(\mu + \nu)\bar{v}$, is constant.

$\Rightarrow \qquad W_i = [(v_i e - c) + e]a_i.$

Alternative expressions for welfare associated with agent i's decision.

Social optimum

$a_i = 1$ iff $v_i \geq v^{SO}$, with $v^{SO}e - c + e = 0$

Equilibrium hinges on reputational incentives

Let $\Delta(v^*) \equiv M^+(v^*) - M^-(v^*) \equiv E[v|v \geq v^*] - E[v|v < v^*]$.

Social norm (SC) iff $\Delta' < 0$. Equilibrium: $a_i = 1$ iff $v_i \geq v^*$

$$v^* e - c + \theta \Delta(v^*) = 0$$

where $\theta = \begin{cases} \mu & \text{(silo information)} \\ \mu + \nu & \text{(social score)} \end{cases}$

▸ over/under provision

III. LEVERAGING SOCIAL SANCTIONS TO CONSOLIDATE POLITICAL POWER

Abstract from features usually associated with an Orwellian state (brutality, misinformation…). Only instrument of state: flow of information.

Agent i takes two actions

- pro- or anti-social action $a_i \in \{0, 1\}$, as earlier. Type v_i (drawn from $F(.)$) measures extent of internalization.
- pro- or anti-government action $b_i \in \{0, 1\}$. Personal cost of toeing the line (benefit if negative) is $\theta_i \sim G(\cdot)$. For simplicity, v_i and θ_i independent (see later).

Agent i's image concerns are only on $\hat{v}_i$ (can be relaxed too).

PAYOFFS

Assumption: b_i observed only by state (not crucial: see later)
Individual's objective function:

$$u_i = (v_i e - c)a_i + \mu \hat{v}_i(I_{ij}) + \nu \hat{v}_i(I_i) - \theta_i b_i$$

Government's objective function:

$$W_g = W_i + \gamma E[b_i]$$

$\gamma \geq 0$: autocratic parameter.

UNBUNDLING

State releases behavior in two realms. Agent i picks

$$b_i = 1 \quad \text{iff} \quad \theta_i \leq 0$$
$$a_i = 1 \quad \text{iff} \quad v_i \geq v_u^*$$

where

$$v_u^* e - c + (\mu + \nu)\Delta(v_u^*) = 0$$

BUNDLING

Suppose state conditions good rating not only a good social behavior, but also on toeing the line:

$$\text{rating} = \begin{cases} 1, & \text{with associated reputation } \hat{v}_1, \text{ if } a_i = 1 \text{ and } b_i = 1 \\ 0, & \text{with associated reputation } \hat{v}_0, \text{ otherwise.} \end{cases}$$

We consider, sequentially

- strong ties society
- weak ties society.

From now on, assume that $c \geq e$ (image concerns are needed to generate pro-social behavior) to shorten analysis.

STRONG TIES ($\mu > 0, \nu = 0$)

Proposition *(ineffectiveness of bundling in a tight knit society).*

When relationships are sustained ($\mu > 0 = \nu$), the state cannot leverage a monopoly position on social ratings in order to consolidate political power. There exists an equilibrium whose outcome is the same as when there is no social rating.

Intuition: information that agents have about each other acts as a counterweight for the information supplied by the state.

WEAK TIES ($\mu = 0, \nu > 0$)

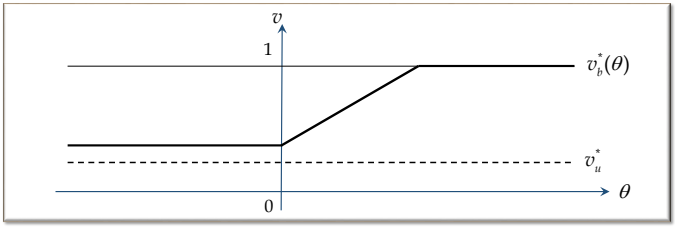
Proposition *(bundling under transient relationships).*

Consider a society with transient relationships ($\mu = 0 < \nu$) *and assume that* $\Delta' \leq 0$. *Under bundling, there exists an equilibrium satisfying:*

(i) *Equilibrium behavior is given by* $v_b^*(\theta) > v_u^*$. *All types* θ *behave less prosocially.*

(ii) *The social contribution* $\bar{a}(\theta) \equiv 1 - F(v_b^*(\theta))$ *is decreasing in* θ.

(iii) *There exists* $\gamma^* > 0$ *such that the government chooses to bundle if and only if* $\gamma \geq \gamma^*$.



Discussion

(a) *Need for monopoly issuance of social ratings*

(b) *Commitment/transparency about way social score is computed*
Algorithm must be transparent; or agents must learn it (information design).

If no commitment: outcome is the same as in the absence of social score.

(c) *Observable b_i*

- no change if $\theta_i \geq 0$ for all i
- negative θ_i types in search for an excuse for not engaging in prosocial acts.

REINTERPRETATIONS

(1) *Divisive issues in a democracy*

In a democracy, $b_i = 0$ stands for action/lifestyle that is disapproved by majority.

- Same logic: majority may want minority to comply (lifestyle, religion, politics…)
- Exact treatment depends on level of interaction between majority and minority, but similar insights.

(2) *Corporate political clout*

Reinterpretation: Platform "rates" official (selective disclosure of facts)

Official i selects two actions

- $a_i \in \{0, 1\}$ affects the welfare of citizens (intrinsic motivation $v_i e - c$). Reputation (re-election) concerns $\nu \hat{v}_i$
- $b_i \in \{0, 1\}$ reflects attitude toward platform (antitrust, tax, legislation on editorial responsibility, etc.). $\theta_i =$ cost of kowtowing to the platform (distribution $G(.)$).

CORRELATED TYPES

Linear-quadratic, Gaussian version of model:

$$u_i = \left[v_i a - \theta_i b - \left(\frac{a^2 + b^2}{2} \right) \right] + \nu \hat{v}(s) [+ \xi \hat{\theta}(s)]$$

and

$$\begin{pmatrix} v_i \\ \theta_i \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \bar{v} \\ \bar{\theta} \end{pmatrix}, \begin{bmatrix} \sigma_v^2 & \rho \sigma_v \sigma_\theta \\ \rho \sigma_v \sigma_\theta & \sigma_\theta^2 \end{bmatrix} \right)$$

Assume a linear social score:

$$s = \alpha a + \beta b$$

At the ruler optimum (ruler maximizes $E[b_i]$):

$$\frac{\beta}{\alpha} = \frac{\sigma_v}{\sigma_\theta} \left[\frac{1}{\rho + \sqrt{1 - \rho^2}} \right] \text{ and } E[a] = \bar{v} + \frac{\nu}{2} \left[1 + \frac{\rho}{\sqrt{1 - \rho^2}} \right]$$

$$\text{and } E[b] = -\bar{\theta} + \frac{\nu}{2} \frac{\sigma_v}{\sigma_\theta} \left(\frac{1}{\sqrt{1 - \rho^2}} \right).$$

Conclusions of linear-quadratic Gaussian case

- State generically benefits from bundling ($\beta \neq 0$)
- Compliance ($E[b]$) increases with
  - image concerns (ν)
  - increase in relative type heterogeneity $\frac{\sigma_v}{\sigma_\theta}$
  - absolute correlation ($|\rho|$)
- Weight β
  - positive if $\rho > -1/\sqrt{2}$
  - negative if $\rho < -1/\sqrt{2}$
- $W_u > W_b$ if (assuming under-signaling: $e > \nu$)
  - either $\rho < 1/\sqrt{2}$ or close to 1
  - e not too large.

NON-IMAGE SANCTIONS (ECONOMIC PENALTIES)

- Even some "economic" sanctions are in part image-driven (blacklisted individuals cannot take first class train or plane)
- Arbitrage $\Rightarrow$ economic sanctions mostly on nominative goods (mobility, visa…)

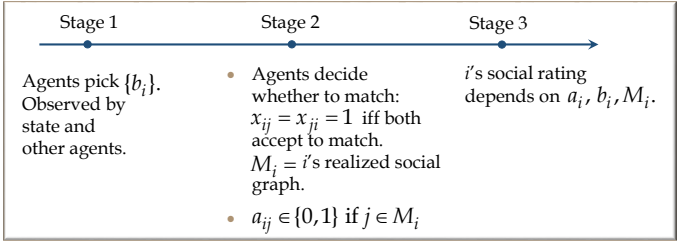
Modeling: DWL. Characterization of behavior under bundling is the same.

IV. GUILT BY ASSOCIATION

- Coloring of a person's perception by the company she keeps: traditional policy in totalitarian regimes.
- Face recognition, AI…has substantially reduced the cost for the state of drawing our exact social graph.

Sequential moves variant of previous model. Agent i picks action $a_{ij} \in \{0,1\}$ for each matched partner j.

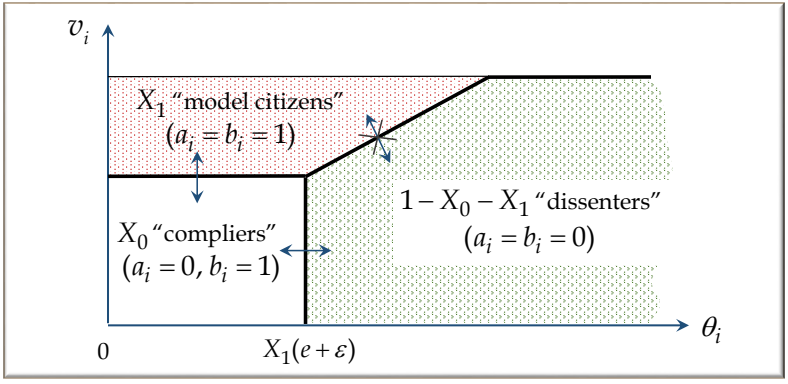
Agents $i, j \in [0,1]$ (mass 1). $v_{ij} = v_i$ for all j. Equilibrium in which $a_{ij} = a_i$ same for all $j \in M_i$.



- Guilt by association: good social rating requires
  - $a_i = b_i = 1$
  - and there exists no j in M_i such that $b_i = 0$.
- Payoff function of individual i

$$u_i = \int_j x_{ij}[(v_i e - c)a_{ij} + e a_{ji} + \varepsilon]dj - \theta_i b_i + \nu \hat{v}_i,$$

  - $\varepsilon \geq 0$ fixed benefit of interaction (irrelevant when M_i exogenous)
  - transient relationships here.
- Assume for simplicity $\theta_i \geq 0$ for all i.
- Rule out WDS/coordination problems (matches occur if both so desire). Still may have multiple equilibria (e.g. no one dissents and all dissent).



Proposition *(social graph).*

Guilt by association makes high-score agents ostracize low-score ones. Incorporating the social graph into the social score is more appealing to an autocratic ruler relative to unbundling. *[Need condition for comparison with simple bundling.]*

Buyer/seller application

"Guilt by association" may apply to commerce as well. Examples of such "ostracism":

- Airline company does not sell business-class ticket, or sells no ticket at all, to individual with low rating
- Talented pupil is not accepted in school because of parents' poor social rating.

Same model, with two sides (buyers, sellers). Sellers themselves are rated (by looking at whom they transact with). New feature: sellers rely on government for authorizations, subsidies…

V. CONCLUDING REMARKS

Social ratings

- have the potential to enhance trust and lead to a more harmonious society
- but may be seriously dysfunctional.

Key challenge for future work: design principle-based policy frameworks.

A few hints

- leave aside information about divisive issues and about the social graph
- beware exclusive or dominant design of social ratings by the state, and political coverage by private platforms.

Even so, many challenging issues, including

- Weights on various components of a social rating (multi-tasking)?
- Granularity (multi-dimensional v_i)?
- Accounting for rating subjectivity
  - sentiments (collusion/retaliation)
  - prejudices and discrimination
  - differences in taste (is the driver "friendly" or "talkative"? Is the restaurant "lively" or "noisy"?)
  - differences in attention paid when rating others.
- Heterogenous image concerns (party member/public sector employee...).
- Should one have a social score in a first place?

THANK YOU FOR YOUR ATTENTION

Target Slides

WHAT CRITERION COULD A SOCIAL SCORE INCLUDE?

- *Some "reasonable"*: credit history, tax compliance, environmentally friendly behavior, traffic violations…
- *Some for which "devil is in the detail"*: spreading of fake news (who decides what is fake news?)…
- *Some definitely "unappealing" ones*: social graph, personal traits, political or religious opinions…

return

RELATED LITERATURES

1) *Economics of privacy*: data collection and analysis enable more effective (but possibly socially detrimental)
   - second- and third-degree price discrimination
   - search and matching
2) *Information design*: commitment to a disclosure rule.
3) *Community enforcement*

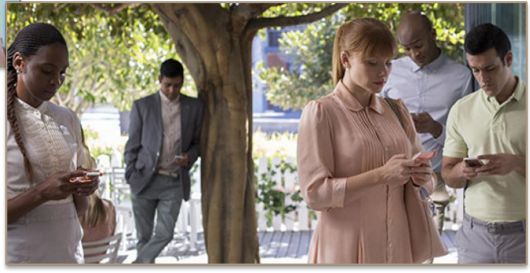
Literature emphasizes benefits from community enforcement and accordingly focuses on equilibria that exhibit a high level of enforcement. In this paper:

- platforms and governments can employ data integration to further their own goals.
- dysfunctional features of such enforcement emerge.

4) *Theoretical and empirical literatures on image concerns*

return

Over signaling



return

TOO MUCH OR TOO LITTLE PROSOCIAL BEHAVIOR?

Social optimum: $v^{SO}e - c + e = 0.$

Equilibrium behavior: If $\Delta' < 0$ (social norm), must assume that image concerns are not too strong if one wants to guarantee uniqueness of cutoff v^* (such that $a_i = 1$ iff $v_i \geq v^*$).

Silo reputations/data islands/privacy

$$v^*e - c + \mu\Delta(v^*) = 0.$$

Underprovision iff $e > e^s \equiv \mu\Delta(v^*)$.

Social score/transparency

$$v^*e - c + (\mu + \nu)\Delta(v^*) = 0.$$

Underprovision iff $e > e^t \equiv (\mu + \nu)\Delta(v^*)$.

▸ Trivia question

return