# Replication, Replication and Replication

## – *Some Hard Lessons from Model Alignment*


http://cfpm.org

**David Hales**

**Centre for Policy Modelling,**

**Manchester Metropolitan University, UK.**

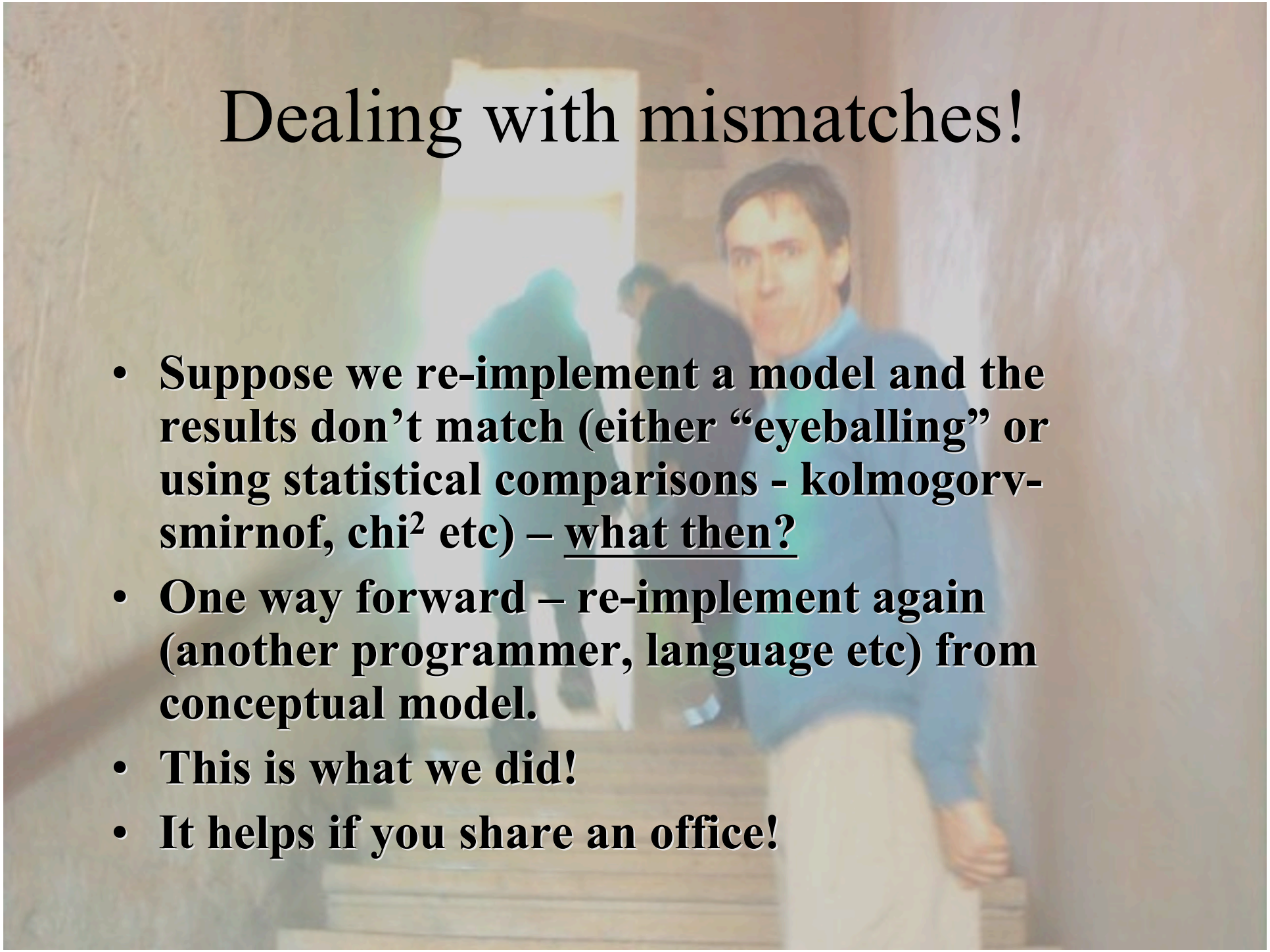
**http://www.davidhales.com**

*M2M Workshop March 31st-1st April 2003.*

# Why replicate?

- **Ensure that we fully understand the conceptual model (as described in a paper)**
- **Check that the published results are correct**
- **Add credibility to the published results (different languages, random number generators, implementations of the conceptual model)**
- **A base-line for further experimentation**

# Dealing with mismatches!



- **Suppose we re-implement a model and the results don't match (either "eyeballing" or using statistical comparisons - kolmogorv-smirnof, chi$^2$ etc) – what then?**
- **One way forward – re-implement again (another programmer, language etc) from conceptual model.**
- **This is what we did!**
- **It helps if you share an office!**

# The model (Riolo et al 2001)

- Tag based model of altruism
- Holland (1992) discussed tags as a powerful "symmetry breaking" mechanism which could be useful for understanding complex "social-like" processes
- Tags are observable labels or social cues
- Agents can observe the tags of others
- Tags evolve in the same way that behavioural traits evolve (mimicry, mutation etc)
- Agents may evolve behavioural traits that discriminate based on tags

# The Model – Riolo et al 2001

- **100 agents, each agent has a tag (real number) and a tolerance (real number)**
- **In each cycle each agent is paired some number of times with a random partner.**
- **If their tags are similar enough (difference is less than or equal to the tolerance) then the agent makes a donation.**
- **Donation involves the giving agent losing fitness (the cost = 0.1) and the receiving agent gaining some fitness (=1)**
- **After each cycle a *tournament selection* process based on fitness, increases the number of copies of successful agents (high fitness) over those with low fitness.**
- **When successful agents are copied, mutation is applied to both tag, tolerance.**

# Original results (Riolo et al)

| Effect of pairings on donation rate | | |
|---|---|---|
| **Parings** | **Donation rate (%)** | **Average tolerance** |
| **1** | **2.1** | **0.009** |
| **2** | **4.3** | **0.007** |
| **3** | **73.6** | **0.019** |
| **4** | **76.8** | **0.021** |
| **6** | **78.7** | **0.024** |
| **8** | **79.2** | **0.025** |
| **10** | **79.2** | **0.024** |

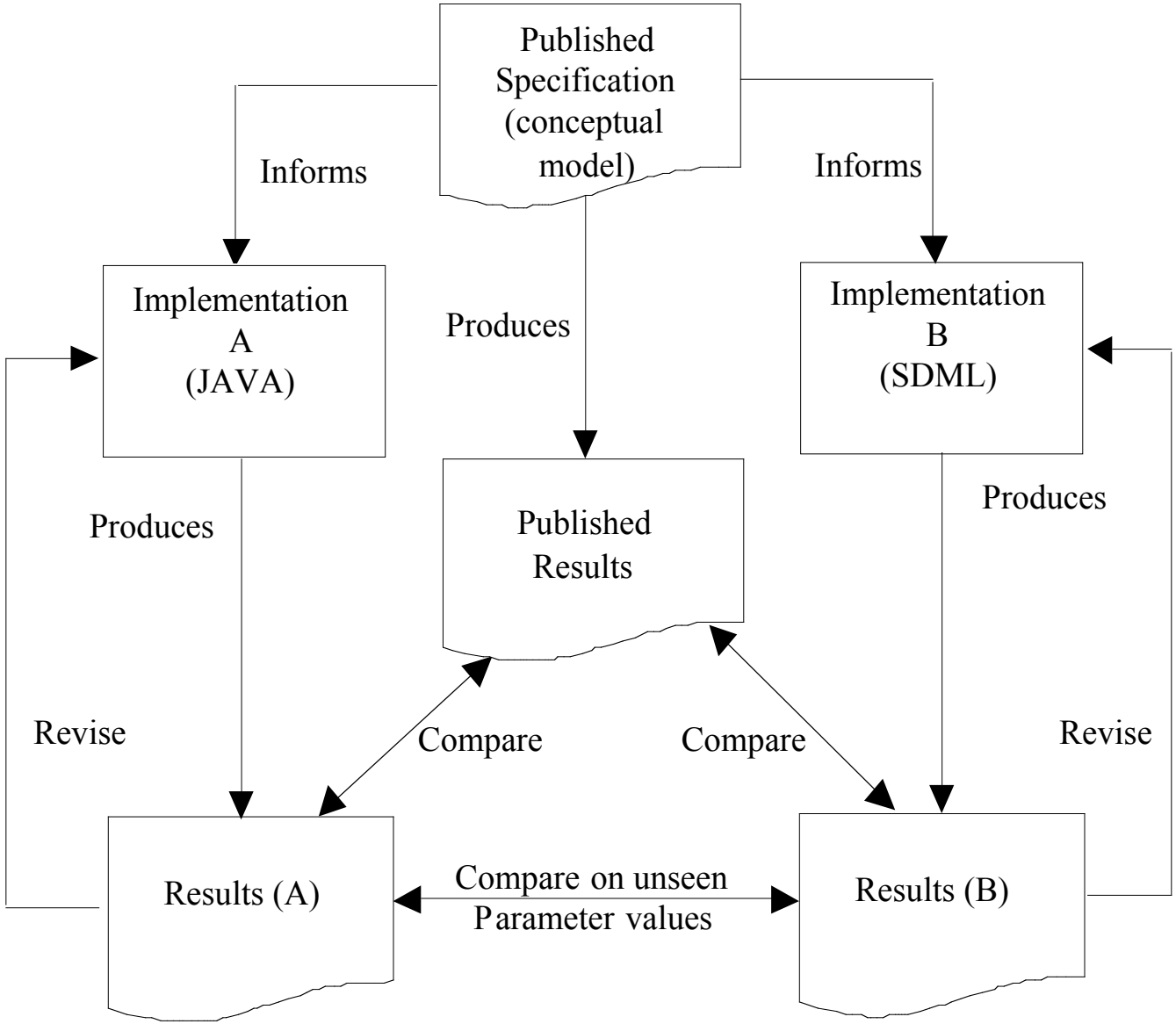
# First re-implementation (A)

| Effect of pairings on donation rate | | |
|---|---|---|
| **Parings** | **Donation rate (%)** | **Average tolerance** |
| **1** | **5.1 (3.0)** | **0.010 (0.1)** |
| **2** | **42.6 (38.3)** | **0.012 (0.5)** |
| **3** | **73.7 (0.1)** | **0.018 (0.1)** |
| **4** | **76.8 (0.0)** | **0.021 (0.0)** |
| **6** | **78.6 (0.1)** | **0.023 (0.1)** |
| **8** | **79.2 (0.0)** | **0.025 (0.0)** |
| **10** | **79.4 (0.2)** | **0.026 (0.2)** |

# Second Re-implementation (B)

- A second implementation reproduced what was produced in the first implementation (A) but not the original results
- Outputs from A and B were checked over a wide range of the parameter space – different costs, agents and awards etc. and they matched.

**Relationship of published model and two re-implementations**



## Re-implementations match but different from original published results

possible sources of the inconsistency:

- Implementation used to produce the published results did not match the published conceptual model.
- Some aspect of the conceptual model was not clearly stated in the published article
- Both re-implementations had somehow been independently and incorrectly implemented (in the same way).

# Three variants of tournament selection

A problem of interpretation was identified in the tournament selection procedure for reproduction. In the original paper it is described thus:

*"After all agents have participated in all parings in a generation agents are reproduced on the basis of their score relative to others. The least fit, median fit, and most fit agents have respectively 0, 1 and 2 as the expected number of their offspring. This is accomplished by comparing each agent with another randomly chosen agent, and giving an offspring to the one with the higher score."*

# Three variants of tournament selection

In both re-implementations the we assumed that when compared agents have *identical scores* a random choice is made between them to decide which to reproduce into the next generation (this is unspecified in the text).

Consequently, there are actually three possibilities for the tournament selection that are consistent with the description in text

# Three Variants Of Tournament Selection

```
LOOP for each agent in population
  Select current agent (a) from pop
  Select random agent (b) from pop
  IF score (a) > score (b) THEN
     Reproduce (a) in next generation
  ELSE IF score (a) < score (b) THEN
     Reproduce (b) in next generation
  ELSE (a) and (b) are equal
     Select randomly (a) or (b) to be
     reproduced into next generation.
  END IF
END LOOP
```

## a) No Bias

```
LOOP for each agent in population
  Select current agent (a) from pop
  Select random agent (b) from pop
  IF score (a) >= score (b) THEN
     Reproduce (a) in next generation
  ELSE score (a) < score (b)
     Reproduce (b) in next generation
  END IF
END LOOP
```

## b) Selected Bias

```
LOOP for each agent in population
  Select current agent (a) from pop
  Select random agent (b) from pop
  IF score (a) <= score (b) THEN
     Reproduce (b) in next generation
  ELSE score (a) > score (b)
     Reproduce (a) in next generation
  END IF
END LOOP
```

## c) Random Bias

# Results from 3 variants

| | **Results From The Three Variants Of Tournament Selection** | | | | | |
|---|---|---|---|---|---|---|
| | No Bias (a) | | Selected Bias (b) | | Random Bias (c) | |
| Parings | Don | Ave. Tol | Don | Ave Tol | Don | Ave Tol |
| 1 | 5.1 | 0.010 | 2.1 | 0.009 | 6.0 | 0.010 |
| 2 | 42.6 | 0.012 | 4.4 | 0.007 | 49.6 | 0.013 |
| 3 | 73.7 | 0.018 | 73.7 | 0.019 | 73.7 | 0.018 |
| 4 | 76.8 | 0.021 | 76.9 | 0.021 | 76.8 | 0.021 |
| 6 | 78.6 | 0.023 | 78.6 | 0.023 | 78.7 | 0.023 |
| 8 | 79.2 | 0.025 | 79.2 | 0.025 | 79.2 | 0.025 |
| 10 | 79.4 | 0.026 | 79.4 | 0.026 | 79.4 | 0.026 |

# Further experimentation

- Now we had two independent implementations of the Riolo model that matched the published results
- We were ready to experiment with the model to explore its robustness
- We changed the model such that donation only occurred if tag values were *strictly less* than the tolerance (we replaced a < with a <= in the comparison for a “tag match”).

# Strictly *less than* tolerance

| **Effect of pairings on donation rate (strict tolerance)** | | |
|---|---|---|
| Parings | Donation rate (%) | Average tolerance |
| 1 | 0.0 | 0.000 |
| 2 | 0.0 | 0.000 |
| 3 | 0.0 | 0.000 |
| 4 | 0.0 | 0.000 |
| 6 | 0.0 | 0.000 |
| 8 | 0.0 | 0.000 |
| 10 | 0.0 | 0.000 |

# Tolerance always set to zero (turned off)

| | **Results when tolerance set to zero for different Pairings** | | | | | |
|---|---|---|---|---|---|---|
| | No Bias (a) | | Selected Bias (b) | | Random Bias (c) | |
| Parings | Don | Ave. Tol | Don | Ave Tol | Don | Ave Tol |
| 1 | 3.1 | 0.000 | 0.0 | 0.000 | 4.1 | 0.000 |
| 2 | 65.4 | 0.000 | 0.0 | 0.000 | 65.6 | 0.000 |
| 3 | 75.3 | 0.000 | 0.0 | 0.000 | 75.4 | 0.000 |
| 4 | 77.6 | 0.000 | 0.0 | 0.000 | 77.7 | 0.000 |
| 6 | 78.8 | 0.000 | 0.0 | 0.000 | 78.8 | 0.000 |
| 8 | 78.9 | 0.000 | 1.9 | 0.000 | 78.9 | 0.000 |
| 10 | 79.0 | 0.000 | 7.6 | 0.000 | 79.0 | 0.000 |

# With noise added to tags (Gaussian zero mean and stdev $10^{-6}$)

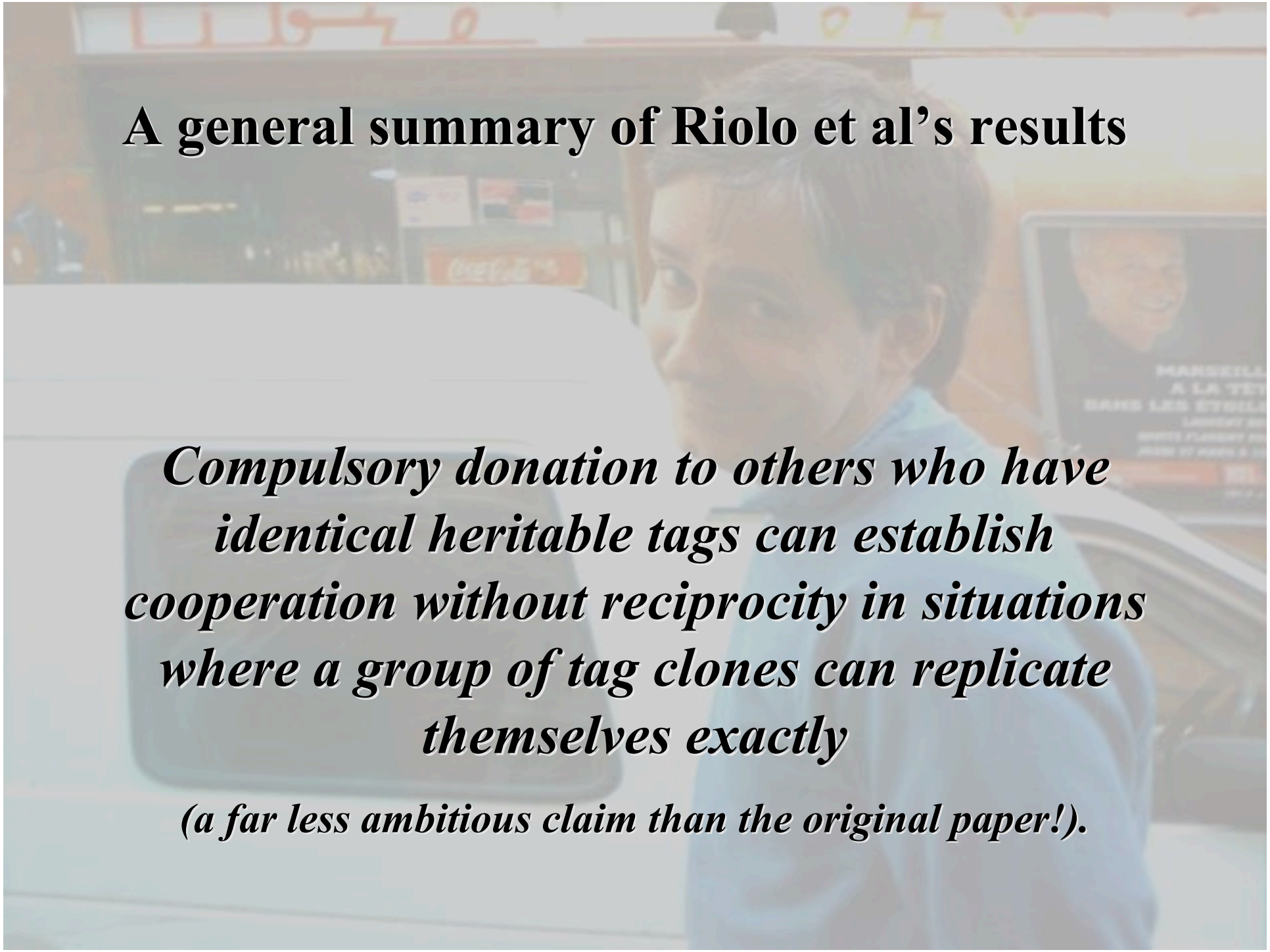
| | **Results when noised added to tag values on reproduction** | | | | | |
|---|---|---|---|---|---|---|
| | No Bias (a) | | Selected Bias (b) | | Random Bias (c) | |
| Parings | Don | Ave. Tol | Don | Ave Tol | Don | Ave Tol |
| 1 | 3.7 | 0.009 | 1.9 | 0.009 | 4.2 | 0.009 |
| 2 | 3.1 | 0.007 | 1.5 | 0.006 | 3.7 | 0.007 |
| 3 | 4.0 | 0.005 | 1.5 | 0.005 | 5.1 | 0.005 |
| 4 | 6.8 | 0.005 | 2.0 | 0.005 | 8.5 | 0.005 |
| 6 | 13.1 | 0.004 | 6.2 | 0.004 | 14.2 | 0.004 |
| 8 | 15.5 | 0.004 | 12.7 | 0.004 | 16.2 | 0.004 |
| 10 | 12.1 | 0.002 | 10.9 | 0.003 | 12.8 | 0.003 |

# What's going on?

- In the "selected bias" setting, with zero tolerance, donation can only occur between "tag clones"
- Since initially it is unlikely that tag clones exist in the population, there is no donation
- The "selected bias" method of reproduction reproduces exactly the same population when all fitness values are equal (zero in this case).
- The other reproduction methods allow for some "noise" in the copying to the next generation such that a tag may be duplicated to two agents in the next generation – even though all fitness scores are the same.
- A superficial analysis might conclude that tolerance *was* important in producing donation – the original paper implies donation based on tolerance. This appears to be *false.*

# A major conclusion

- *The multiple re-implementations gave deeper insight into important (and previously hidden) aspects of the model.*
- *These have implications with respect to possible interpretations of the results.*
- *Confident of critique due to multiple implementations behaving in the same way and aligning with original results.*



# A general summary of Riolo et al's results

***Compulsory donation to others who have identical heritable tags can establish cooperation without reciprocity in situations where a group of tag clones can replicate themselves exactly***

***(a far less ambitious claim than the original paper!).***

# Some practical lessons about aligning models

- Compare simulations first time cycles  checks if initialised, same clearer than after new effects emerge or chaos appears
- Use statistical tests over long-term averages of many runs, (eg. Kolmogorov-Smirnov) to test if figures come from the same distribution
- When simulations don't align, progressively turn off features of the simulation (e.g. donation, reproduction, mutation etc.) until they do align. Then progressively reintroduce the features.
- Use different kinds of languages to re-implement a simulation (we used a declarative and an imperative language) programmed by different people.

# Conclusions / Suggestions

- **Description of a published ABM should be sufficient for others to re-implement**
- **Results should be independently replicated and confirmed before they are taken seriously**
- **Results can not be confirmed as correct but merely survive repeated attempts at refutation**
- **Everyone should get their students to replicate results from the literature (how many will survive? Certainly *not* 100% !)**

# M2M – open issues

- **Cioffi-Revilla – is a "typology" of ABM possible/desirable (dimensions: space/time/theoretical/empirical)?**
- **Kirman – how do we move ABM towards (at least a partial) common framework (what would it look like)?**
- **Janssen – using ABM to understand more clearly what existing learning functions can do – their biases, strengths.**
- **Clarify and test model results via replication – Rouchier, Edmonds, Hales.**
- **Duboz & Edwards – Integrating top-down and bottom-up / virtual experiments – higher-level abstractions. But what about application to more *complex ABS?***
- **Kluver – topology as key dimension in soft comp. algorithms – results converged but how to identify *differences?***
- **Gotts – Trap$^2$ (formalised, abstracted classes of ABS with an interpretation). How to formalise / discover ?**
- **Flache – steps in alignment – can these be generalised (Heuristics)?**


# JASSS Special Issue Timetable

- May 1st – Deadline for new or substantially revised submissions
- July 1st – Deadline for reviews
- August 1st – Deadline for revised papers
- October – JASSS Special Issue

# Goodbye from M2M-1….

**Big thanks to GREQAM for hosting and organising!**

**2005 – M2M2 ?** **(http://cfpm.org/m2m).**

***(provisionally: Nick Gotts, Claudio Cioffi-Revilla, Guillaume Deffaunt)***

**MABS2003 @ AAMAS2003 (Melbourne July) (http://cfpm.org/mabs2003).**

**"Frontiers of Agent Based Social Simulation" (Kluwer) – call for chapters soon – attempt to address some of the open issues identified here.**