Intervening in science policy (The debate on reproducibility)

Sabina Leonelli
Exeter Centre for the Study of Life Sciences (Egenis),
University of Exeter
@sabinaleonelli

Outline

- Framing STS interventions
- *CASE: Reproducibility debate*
- How to extract 'lessons' from STS analysis?
- *CASE: my own policy work on reproducibility*
- Rounding up the summer school: ways forward for Open Science

STS interventions: key issues

- Who is contributing to what?
  - Crucial not just to 'make a difference', but to understand the field (two-way street with asymmetric twist: we learn more than we contribute)
  - Concentric circles: intervening starts at home
- What to contribute?
  - Understanding stakes and motivations is critical for intervention
  - Big critiques and distinctions are critical to advance our knowledge and vision, but not always productive when presented to other publics
  - Beware of polarizing effects, pay attention to low-hanging fruits and existing openings

STS interventions: key issues

- For which goals and ideas of common good?
  - Interventions are extremely fraught: our research teaches us there's no ideal / one-sided solution.
  - What claims do we have over truth / 'public' benefits? What are our stakes, and why? In whose interest are we operating? Who are are peers and our publics?
- What do we get out of it?
  - Issues with rewards and recognition within STS itself: vis-à-vis academic and non-academic jobs
  - Paves the way for life outside of academia rather than life inside it!?

Discussion

- What makes us accountable, to whom?
- Should our research be reproducible, how and for which purposes?

Reproducibility revisited

The reproducibility "crisis"

- Are methods failing?
  - Questionable uses of statistical techniques to smoothen bias and exclude uncomfortable results (e.g. p-hacking, selective reporting)
  - Confusion around scale of data analysis and trustworthiness of data sources / processing
  - Ineffectual quality control & lack of clarity around who is responsible
  - Widespread mistrust of published results

- Pursuit of reproducibility as overarching epistemic value
  - the extent to which consistent results are obtained when an experiment is repeated

(Collins 1985, Radder 1996, Romero 2019, Feest 2019, Guttinger 2020, Mayo 2020, Machery 2020)

the reproducibility "crisis"

Reproducibility comes in a variety of forms geared to different methods, settings, targets and goals in science:

- Assumed degree of **control** over research conditions
  - choice of variables vs what can/should be stabilized
- Understanding of **variation**
  - phenomenon to be explained, confounder or signal of error?
- Dependence on **statistics** and **computation**
  - as inferential tools
- Precision of the research **goals**
  - from exploratory research to hypothesis testing)
- Stability of **background knowledge** and evidence base
- Dependence on researchers' **judgment**
  - role of expertise and related training

Leonelli, S (2018) Re-Thinking Reproducibility as a Criterion for Research Quality. *Research in the History of Economic Thought and Methodology* 36B, 129-146

Type of Reproducibility	Assumed control	Dependence on statistics	Precision of goals	Dependence on judgement
Computational Reproducibility	total	high	high	none
Direct Experimental Reproducibility	high	high	high	low
Scoping/Indirect/Hypothetical Reprod.	limited	variable	limited	variable
Reproducible Expertise	variable	variable	variable	high
Reproducible Observation	low	low	low	high
Irreproducible Research	none	low	low	total

Leonelli, S (2018) Re-Thinking Reproducibility as a Criterion for Research Quality. *Research in the History of Economic Thought and Methodology* 36B, 129-146

Type of Reproducibility	Assumed control	Dependence on statistics	Precision of goals	Dependence on judgement
Computational Reproducibility	total	high	high	none
Direct Experimental Reproducibility	high	high	high	low
ypothetical Reprod.	limited	variable	limited	variable
Reproducible Expertise	ariable	variable	variable	high
Reproducible Observation	low	low	low	high
Irreproducible Research	none	low	low	total

Overly narrow interpretation

- Highly controlled experiments with pre-specified goals exemplify "best practice"..
- .. doing no justice to other research methods
  - e.g. data-intensive discovery and qualitative traditions focused on analysis of situatedness
- False dichotomy of "hermeneutic" and "quantitative" approaches
  - Devalues role of expertise and embodied knowledge in data production, processing and assessment… as well as significance of social context
    - E.g. when a study needs to be completely redesigned in order to be replicated, because social context has changed
  - Does not help to distinguish unintentional mistakes, cheating, difference in research conditions, constructive vs malicious questioning of accepted 'facts'

Overly Broad Interpretation

- Common conceptual confusion:
  - **Generalizability**: scope of research is not the same as quality
  - **Sharing**: making research accessible does not improve quality, as long as there is no scrutiny → what makes a difference is re-use and discrimination around what is worth sharing
  - **Transparency:** Imbalance of requirements for publicly and privately sponsored research
- Reproducibility does not cover all concerns around invisible work:
  - **Scalability and optimization** (e.g. software for clinical work needs to be optimized for large patient pool)
  - **Transdisciplinarity**: set-up of collaborations and initiatives
  - **Translation**: gap in support for efforts towards bringing research to market

Formulating "policy advice"

What are the 'lessons learnt' from an analysis of this kind?

- Relation to scholarly outputs
- Opportunities for understanding
- Accountability for implications

Examplar of slides from my own policy interventions

(Vlanders Thinkers programme)

How does the pursuit of reproducibility help address scientific crisis?

- does not necessarily 'fix' concerns around research quality
- does not provide a universal solution, since reproducibility means different things to different fields/problems/approaches
- does not address systemic issues with rewards and incentives
  - e.g. entrenched hierarchies of credit and expertise

How does the pursuit of reproducibility help address scientific crisis?

- Need to explore systemic reasons for "crisis of reproducibility":
  - Lack of incentives and resources for researchers to explicitly and regularly discuss
    1. methodological commitments within and across disciplines, and beyond academia
    2. how learning from mistakes and problems happens in everyday practice - and is documented
    3. the strategies used to choose which research components need to be preserved in the long term, and how
  - Side-lining of open science and research geared towards community benefit
  - Credit system vis-à-vis early career researchers, technicians and support staff
  - Emphasis on short-term outcomes
    - E.g. Reliance on automation and data-intensive tech to provide a "quick fix"

Getting incentives right: Flanders

- **Rewards and incentives** lagging behind:
  - Still metrics obsessed – emphasis on the short term
  - Even in places where metrics are complemented by qualitative evaluation, funders/university reward novelty over replication/quality
  - Reviewing activities remain invisible work: volunteered, unrewarded
  - Emphasis on transparency for publicly funded researchers, while industry receives no such scrutiny
  - Generalizable results favoured over robust results
- This in turn prompts **conflict of interests and goals:**
  - junior-senior staff, students-supervisors
  - collaborators across institutions and countries
  - disciplines
  - professional staff-academics
  - industry-academia
  - and more generally: How to build a research culture of open discussion, when everybody is monitoring everybody else for signs of 'bad faith'?

Getting incentives right: Flanders

- **Recognition:** needs to come in substantive forms, e.g. hiring and promotion criteria
  - "badges" and prizes can foster "open washing"
- **Funding:**
  - increasing emphasis on integrity ("second axis" of assessment) and negative results (FWO)
  - what counts as "new ideas"?
  - how to support transdisciplinary research?
  - critical role of (international) assessment panels
- **Training:** crucial but does not resolve all issues
  - Not just a matter of 'research culture'
  - Important not to let full weight of R requirement fall on researchers
- **Support**: key role of data steward and integrity officers
  - help with expertise but also to mediate conflict

Discussion - What works and what does not?

A few words to round up..

Ways forward for a pluralistic OS

- recognise danger of overgeneralized principles and standards
  - *including* openness
- cultivate active resistance against entrenched discrimination, including dominance of long-standing repertoires
  - seek to understand where it is coming from
- develop community-specific, value-laden criteria for research quality
  - quality does matter..
- identify and share burdens of OS implementation
  - dissent/critique when required
- foster ongoing debate on what counts as science
  - systematically probing existing boundaries for systems of practice and related governance/institutions

Thank you

European Research Council

Established by the European Commission

The Alan Turing Institute

Wissenschaftskolleg zu Berlin

Practical example: Preregistration

- ✓ A way to formalize and remember the rationale for specific choices at a given moment of the research process

- ✗ Resource-intensive tool for research quality assessment (NOT just a matter of comparing plans and outcomes)

Practical example: Preregistration

- ✓ A way to formalize and remember the rationale for specific choices at a given moment of the research process

- ✗ Resource-intensive tool for research quality assessment (NOT just a matter of comparing plans and outcomes)

Five forms of reproducibility

1. Computational reproducibility
2. Direct experimental reproducibility (highly standardized experiments)
3. Scoping/Indirect/Hypothetical reproducibility (semi-standardized experiments)
4. Reproducible expertise
5. Reproducible observation

Five forms of reproducibility

1. Computational reproducibility
2. Direct experimental reproducibility (highly standardized experiments)
3. Scoping/Indirect/Hypothetical reproducibility (semi-standardized experiments)
4. Reproducible expertise
5. Reproducible observation

1. Computational reproducibility

- Researchers focus on finding and resolving mistakes and bugs in data analysis by running the same data through a given set of algorithms over and over again
- Key condition: open & reusable code and data

Fig. 2. Reproducibility Spectrum According to Peng (2011), Redrawn by Michel Durinx.

1. Computational reproducibility

Assumed degree of control over research conditions	TOTAL
Dependence on statistics as inferential tool	HIGH
Precision of the research goals	HIGH
Dependence on researchers' judgment	LOW

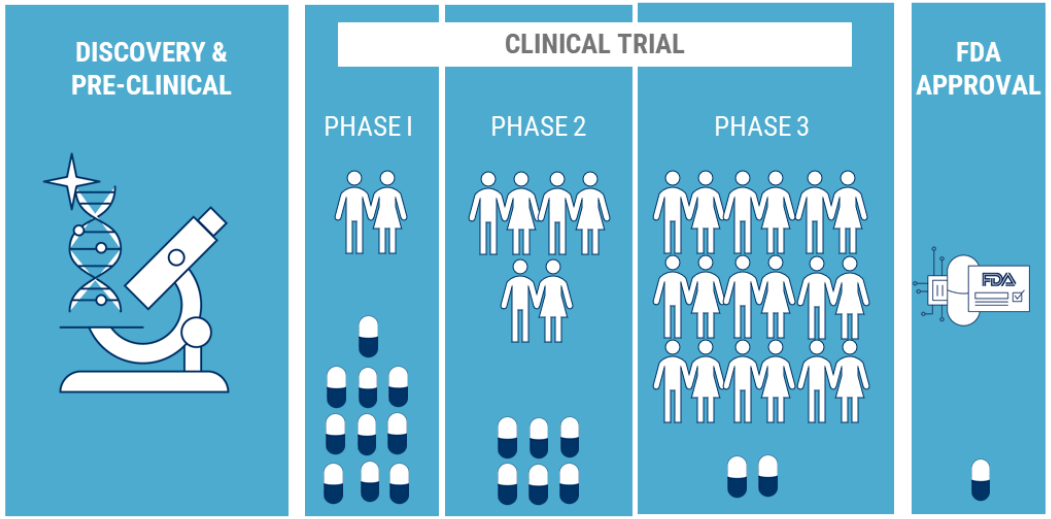
Five forms of reproducibility

1. Computational reproducibility
2. Direct experimental reproducibility (highly standardized experiments)
3. Scoping/Indirect/Hypothetical reproducibility (semi-standardized experiments)
4. Reproducible expertise
5. Reproducible observation

2. Direct experimental reproducibility: standardised experiments

- the ability to obtain the same results through the repeated application of the same research methods/processes

e.g. clinical trials



Source: cbinsights.com

2. Direct experimental reproducibility: standardised experiments

Assumed degree of control over research conditions	HIGH
Dependence on statistics as inferential tool	HIGH
Precision of the research goals	HIGH
Dependence on researchers' judgment	LOW

Five forms of reproducibility

1. Computational reproducibility
2. Direct experimental reproducibility (highly standardized experiments)
3. Scoping/Indirect/Hypothetical reproducibility (semi-standardized experiments)
4. Reproducible expertise
5. Reproducible observation

3. Semi-Standardised Experiments

- methods, set-up and materials construed with ingenuity in order to yield very specific outcomes
- yet some significant parts of the set-up necessarily elude the controls set up by experimenters

E.g. Discovery / pre-clinical research (Lowe, Leonelli and Davies 2019), experiments on model organisms (Ankeny and Leonelli 2020), developmental biology & physiology (Weber, Love)

Psychological experiments on social groups selected because conforming to given physical, social and behavioural criteria, and yet presenting unforeseen sources of variability of potential relevance to the outcomes being generated (Felt 2019)

3. Semi-Standardised Experiments

Assumed degree of control over research conditions	VARIABLE
Dependence on statistics as inferential tool	VARIABLE
Precision of the research goals	LIMITED
Dependence on researchers' judgment	VARIABLE

3a. Scoping Reproducibility

- spot differences in the results obtained by repeating the same experiment
- identify and study sources of variation that may prove significant when interpreting the resulting data

(Leonelli 2018)

3b. Indirect Reproducibility

- obtaining similar results from the performance of *different* experiment
- constitutes a useful validation tool to see whether results produced under variable circumstances converge or not

(Hans Radder)

3c. Hypothetical Reproducibility

- attempt to obtain outcomes that match those *predicted as implications of previous findings*, thereby confirming the reliability of the previous findings

(Felipe Romero 2017)

Five forms of reproducibility

1. Computational reproducibility
2. Direct experimental reproducibility (highly standardized experiments)
3. Scoping/Indirect/Hypothetical reproducibility (semi-standardized experiments)
4. Reproducible expertise
5. Reproducible observation

4. Reproducible Expertise: Non-Standard or Very Expensive Experiments, Research on Rare Materials

- the expectation that any skilled researcher working with the same methods and the same type of materials at that particular time and place would produce similar results
  - e.g. paleontology, archeology, history; highly exploratory research
- Apposite methodologies have been developed to cope with the impossibility to directly replicate the findings
  - vetted access
  - cross-samples research
  - centralisation of research in locations where many researchers can work together, check each other's work and ensure its reliability for those with no access to the same instruments / sources

4. Reproducible Expertise: Non-Standard or Very Expensive Experiments, Research on Rare Materials

Assumed degree of control over research conditions	VARIABLE
Dependence on statistics as inferential tool	VARIABLE
Precision of the research goals	VARIABLE
Dependence on researchers' judgment	HIGH

Five forms of reproducibility

1. Computational reproducibility
2. Direct experimental reproducibility (highly standardized experiments)
3. Scoping/Indirect/Hypothetical reproducibility (semi-standardized experiments)
4. Reproducible expertise
5. Reproducible observation

5. Reproducible Observation: Non-experimental case description

- the expectation that any researcher with similar skills placed in the same time and place would pick out, if not the same data, at least same overarching patterns

e.g. fieldwork in ethology / STS?

Courtesy of University of Pennsylvania

5. Reproducible Observation: Non-experimental case description

Assumed degree of control over research conditions	LOW
Dependence on statistics as inferential tool	LOW
Precision of the research goals	LOW
Dependence on researchers' judgment	HIGH

Overview

Type of Reproducibility	Assumed control	Dependence on statistics	Precision of goals	Dependence on judgement
Computational Reproducibility	******	*****	*****	*
Direct Experimental R	*****	*****	*****	**
Scoping/Indirect /Hypothetical R	***	****	***	****
Reproducible Expertise	****	****	****	*****
Reproducible Observation	**	**	**	*****
Irreproducible Research	*	**	**	******