Human-Centered Computing
Research Group

Freie Universität Berlin

«Human-Centered Data Science»

Enabling Reproducibility of Your Data Science Practice

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

May 12, 2022

Lecture Overview

Recap

Theoretical Considerations of Accountability (aspects of accountability, relationships, professional accountability = reproducibility, definition of repeatability, reproducibility and replicability, degree of replication)

Break

Translating into Practice (recommendations for sharing, using, and understanding your data science workflows, stages of a basic reproducible workflow and things to think of)

Recap

What is Ethics?

"Ethics as a discipline is the philosophical and systematic study of morality which is expressed in enquiries into the nature of the good life, how we should live, what kind of society we want to live in, and how we should treat others."

Ethics is a philosophical discipline.

Morality is a set of moral norms, feelings, attitudes, or actions.

Frauenberger, Christopher, Marjo Rauhala, und Geraldine Fitzpatrick. „In-Action Ethics". _Interacting with Computers_ 29, Nr. 2 (8. März 2017): 220–36.

The Global Landscape of AI Ethics Guidelines

Table 1 | Ethics guidelines for AI by country of issuer (Australia–UK)

Name of document/website	Issuer	Country of issuer
Artificial Intelligence. Australia's Ethics Framework: A Discussion Paper	Department of Industry Innovation and Science	Australia
Montréal Declaration: Responsible AI	Université de Montréal	Canada
Work in the Age of Artificial Intelligence. Four Perspectives on the Economy, Employment, Skills and Ethics	Ministry of Economic Affairs and Employment	Finland
Tieto's AI Ethics Guidelines	Tieto	Finland
Commitments and Principles	OP Group	Finland
How Can Humans Keep the Upper Hand? Report on the Ethical Matters Raised by AI Algorithms	French Data Protection Authority (CNIL)	France
For a Meaningful Artificial Intelligence. Towards a French and European Strategy	Mission Villani	France
Ethique de la Recherche en Robotique	CERNA (Allistene)	France
AI Guidelines	Deutsche Telekom	Germany
SAP's Guiding Principles for Artificial Intelligence	SAP	Germany
Automated and Connected Driving: Report	Federal Ministry of Transport and Digital Infrastructure, Ethics Commission	Germany
Ethics Policy	Icelandic Institute for Intelligent Machines (IIIM)	Iceland
Discussion Paper: National Strategy for Artificial Intelligence	National Institution for Transforming India (NITI Aayog)	India
L'intelligenzia Artificiale al Servizio del Cittadino	Agenzia per l'Italia Digitale (AGID)	Italy
The Japanese Society for Artificial Intelligence Ethical Guidelines	Japanese Society for Artificial Intelligence	Japan
Report on Artificial Intelligence and Human Society (unofficial translation)	Advisory Board on Artificial Intelligence and Human Society (initiative of the Minister of State for Science and Technology Policy)	Japan
Draft AI R&D Guidelines for International Discussions	Institute for Information and Communications Policy (IICP), The Conference toward AI Network Society	Japan
Sony Group AI Ethics Guidelines	Sony	Japan
Human Rights in the Robot Age Report	The Rathenau Institute	Netherlands
Dutch Artificial Intelligence Manifesto	Special Interest Group on Artificial Intelligence (SIGAI), ICT Platform Netherlands (IPN)	Netherlands
Artificial Intelligence and Privacy	The Norwegian Data Protection Authority	Norway
Discussion Paper on Artificial Intelligence (AI) and Personal Data—Fostering Responsible Development and Adoption of AI	Personal Data Protection Commission Singapore	Singapore
Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society	Government of the Republic of Korea	South Korea
AI Principles of Telefónica	Telefónica	Spain
AI Principles & Ethics	Smart Dubai	UAE
Principles of robotics	Engineering and Physical Sciences Research Council UK (EPSRC)	UK
The Ethics of Code: Developing AI for Business with Five Core Principles	Sage	UK
Big Data, Artificial Intelligence, Machine Learning and Data Protection	Information Commissioner's Office	UK
DeepMind Ethics & Society Principles	DeepMind Ethics & Society	UK
Business Ethics and Artificial Intelligence	Institute of Business Ethics	UK
AI in the UK: Ready, Willing and Able?	UK House of Lords, Select Committee on Artificial Intelligence	UK
Artificial Intelligence (AI) in Health	Royal College of Physicians	UK
Initial Code of Conduct for Data-Driven Health and Care Technology	UK Department of Health & Social Care	UK
Ethics Framework: Responsible AI	Machine Intelligence Garage Ethics Committee	UK
The Responsible AI Framework	PricewaterhouseCoopers UK	UK
Responsible AI and Robotics. An Ethical Framework.	Accenture UK	UK
Machine Learning: The Power and Promise of Computers that Learn by Example	The Royal Society	UK
Ethical, Social, and Political Challenges of Artificial Intelligence in Health	Future Advocacy	UK

Table 2 | Ethics guidelines for AI by country of issuer (USA, international, EU and N/A)

Name of document/website	Issuer	Country of issuer
Unified Ethical Frame for Big Data Analysis. IAF Big Data Ethics Initiative, Part A	The Information Accountability Foundation	USA
The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA
Statement on Algorithmic Transparency and Accountability	Association for Computing Machinery (ACM)	USA
AI Principles	Future of Life Institute	USA
AI—Our Approach	Microsoft	USA
Artificial Intelligence. The Public Policy Opportunity	Intel Corporation	USA
IBM's Principles for Trust and Transparency	IBM	USA
OpenAI Charter	OpenAI	USA
Our Principles	Google	USA
Policy Recommendations on Augmented Intelligence in Health Care H-480.940	American Medical Association (AMA)	USA
Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers and Developers	IBM	USA
Governing Artificial Intelligence. Upholding Human Rights & Dignity	Data & Society	USA
Intel's AI Privacy Policy White Paper. Protecting Individuals' Privacy and Data in the Artificial Intelligence World	Intel Corporation	USA
Introducing Unity's Guiding Principles for Ethical AI—Unity Blog	Unity Technologies	USA
Digital Decisions	Center for Democracy & Technology	USA
Science, Law and Society (SLS) Initiative	The Future Society	USA
AI Now 2018 Report	AI Now Institute	USA
Responsible Bots: 10 Guidelines for Developers of Conversational AI	Microsoft	USA
Preparing for the Future of Artificial Intelligence	Executive Office of the President; National Science and Technology Council; Committee on Technology	USA
The National Artificial Intelligence Research and Development Strategic Plan	National Science and Technology Council; Networking and Information Technology Research and Development Subcommittee	USA
AI Now 2017 Report	AI Now Institute	USA
Position on Robotics and Artificial Intelligence	The Greens (Green Working Group Robots)	EU
Report with Recommendations to the Commission on Civil Law Rules on Robotics	European Parliament	EU
Ethics Guidelines for Trustworthy AI	High-Level Expert Group on Artificial Intelligence	EU
AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	AI4People	EU
European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment	Council of Europe: European Commission for the Efficiency of Justice (CEPEJ)	EU
Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems	European Commission, European Group on Ethics in Science and New Technologies	EU
Artificial Intelligence and Machine Learning: Policy Paper	Internet Society	International
Report of COMEST on Robotics Ethics	COMEST/UNESCO	International
Ethical Principles for Artificial Intelligence and Data Analytics	Software & Information Industry Association (SIIA), Public Policy Division	International
ITI AI Policy Principles	Information Technology Industry Council (ITI)	International
Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	International
Top 10 Principles for Ethical Artificial Intelligence	UNI Global Union	International
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	Future of Humanity Institute; University of Oxford; Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI	International
White Paper: How to Prevent Discriminatory Outcomes in Machine Learning	WEF, Global Future Council on Human Rights 2016-2018	International

Continued

Table 2 | Ethics guidelines for AI by country of issuer (USA, international, EU and N/A) (Continued)

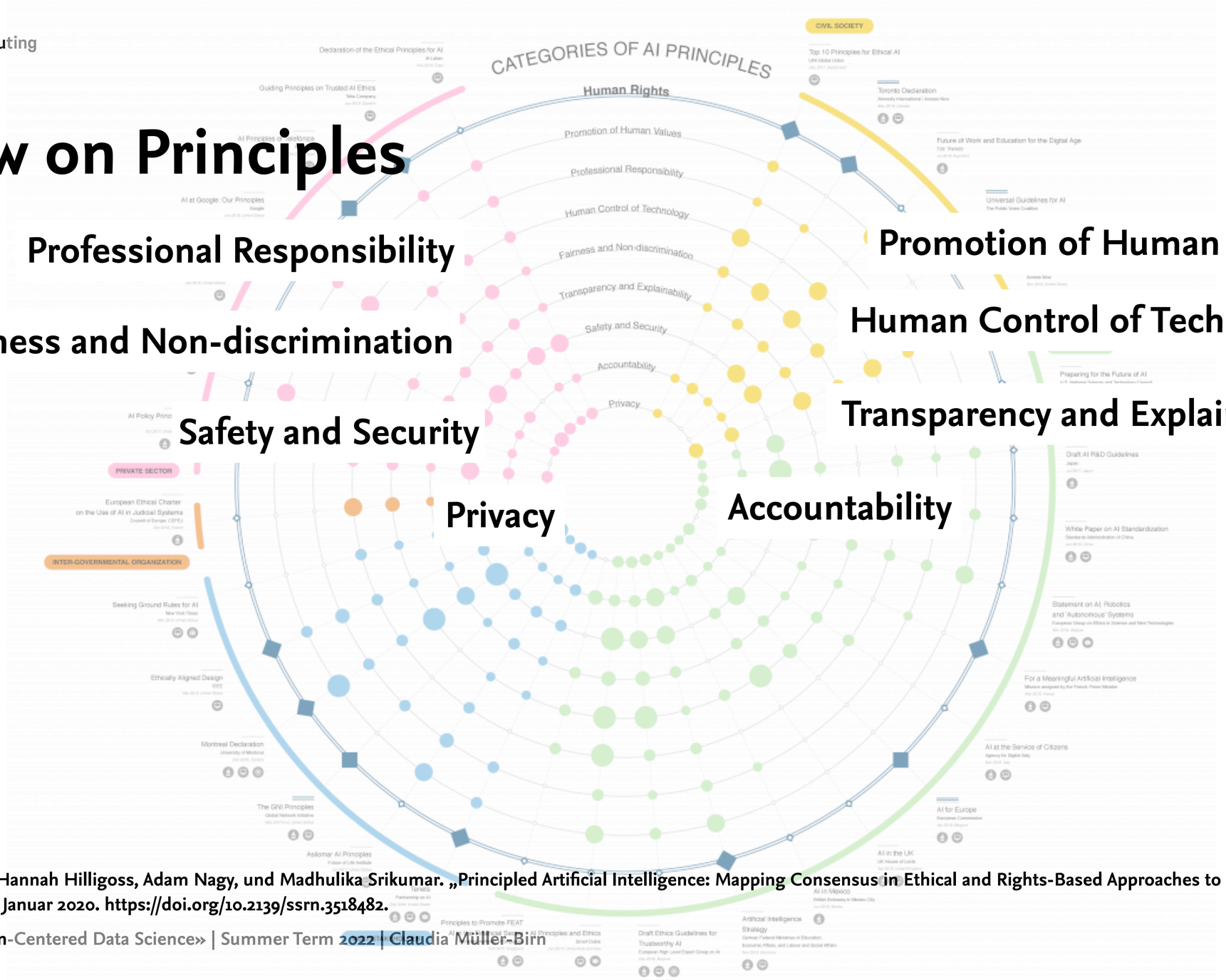
Name of document/website	Issuer	Country of issuer
Privacy and Freedom of Expression in the Age of Artificial Intelligence	Privacy International & Article 19	International
The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems	Access Now; Amnesty International	International
Charlevoix Common Vision for the Future of Artificial Intelligence	Leaders of the G7	International
Artificial Intelligence: Open Questions About Gender Inclusion	W20	International
Declaration on Ethics and Data Protection in Artificial Intelligence	ICDPPC	International
Universal Guidelines for Artificial Intelligence	The Public Voice	International
Ethics of AI in Radiology: European and North American Multisociety Statement	American College of Radiology; European Society of Radiology; Radiology Society of North America; Society for Imaging Informatics in Medicine; European Society of Medical Imaging Informatics; Canadian Association of Radiologists; American Association of Physicists in Medicine	International
Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition (EAD1e)	Institute of Electrical and Electronics Engineers (IEEE), The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems	International
Tenets	Partnership on AI	N/A
Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Fairness, Accountability, and Transparency in Machine Learning (FATML)	N/A
10 Principles of Responsible AI	Women Leading in AI	N/A

84 guidelines

30+ principles

Jobin, Anna, Marcello Ienca, und Effy Vayena. „The global landscape of AI ethics guidelines". Nature Machine Intelligence 1, Nr. 9 (1. September 2019): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Overview on Principles



Professional Responsibility

Promotion of Human Values

Fairness and Non-discrimination

Human Control of Technology

Safety and Security

Transparency and Explainability

Privacy

Accountability

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, und Madhulika Srikumar. „Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI“. SSRN Electronic Journal, 15. Januar 2020. https://doi.org/10.2139/ssrn.3518482.

Incorporating Principles in Your Data Science Practice

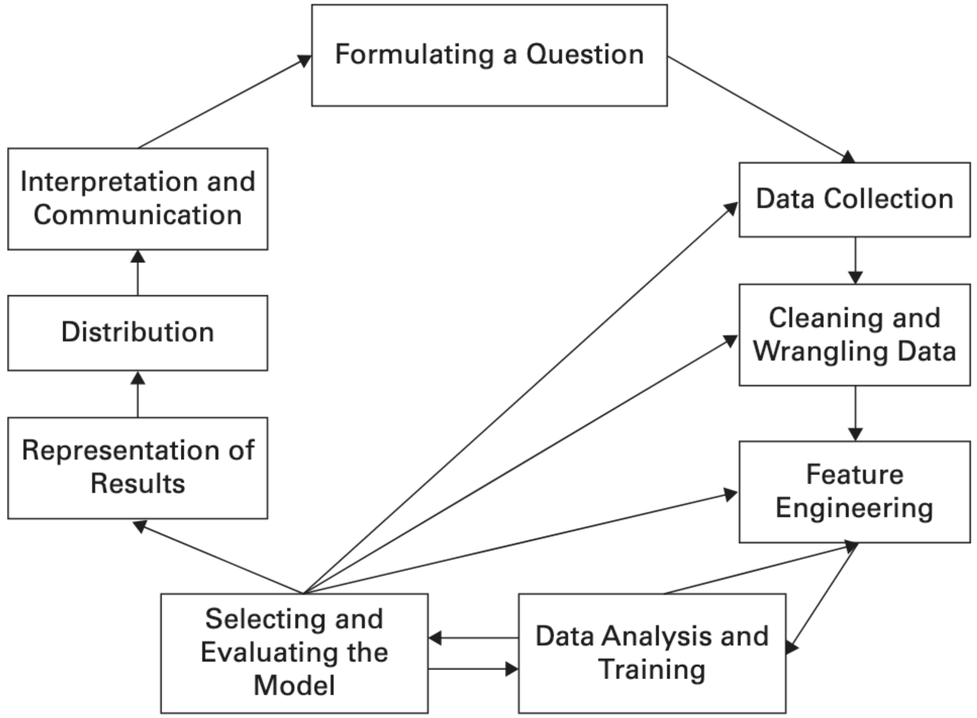
Principles

- Professional Responsibility
- Promotion of Human Values
- Fairness and Non-discrimination
- Human Control of Technology
- Safety and Security
- Transparency and Explainability
- Accountability
- Privacy

Values

Development of a shared ethos

Values Levers

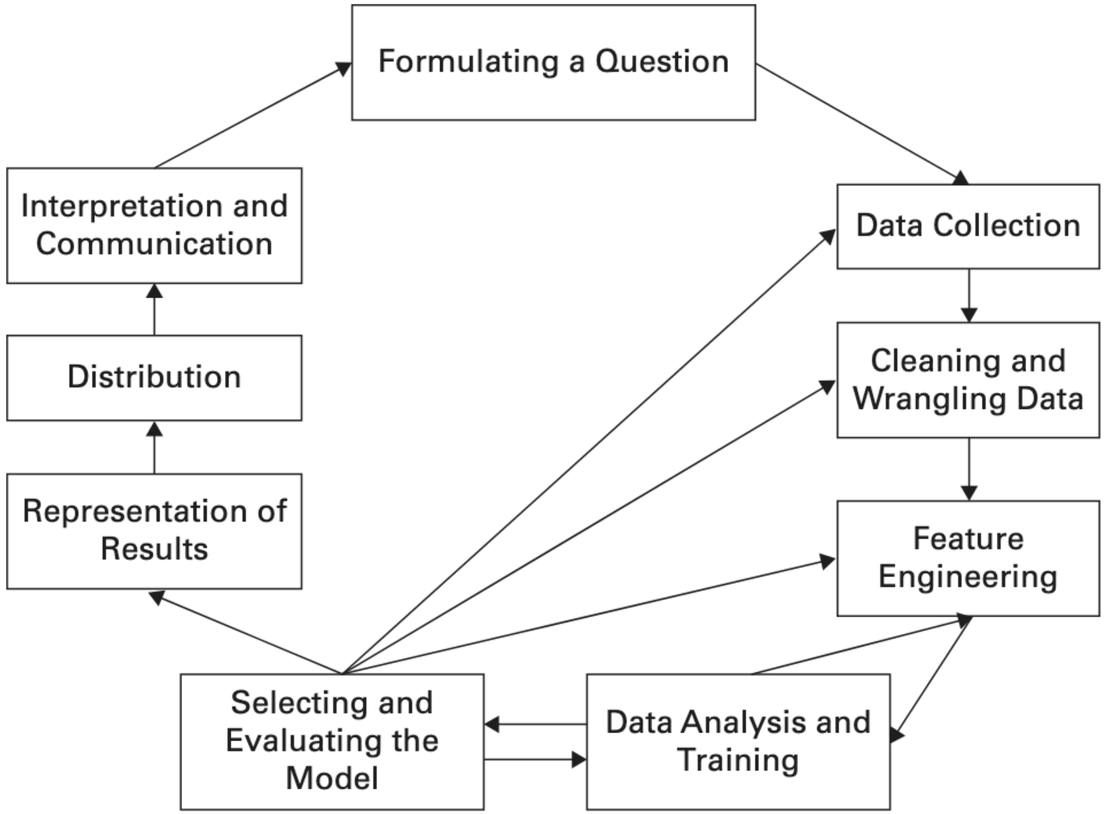


Shilton, Katie. „Values and Ethics in Human-Computer Interaction“. _Foundations and Trends® Human–Computer Interaction_ 12, Nr. 2 (16. Juli 2018): 107–71.

Incorporating Principles in Your Data Science Practice

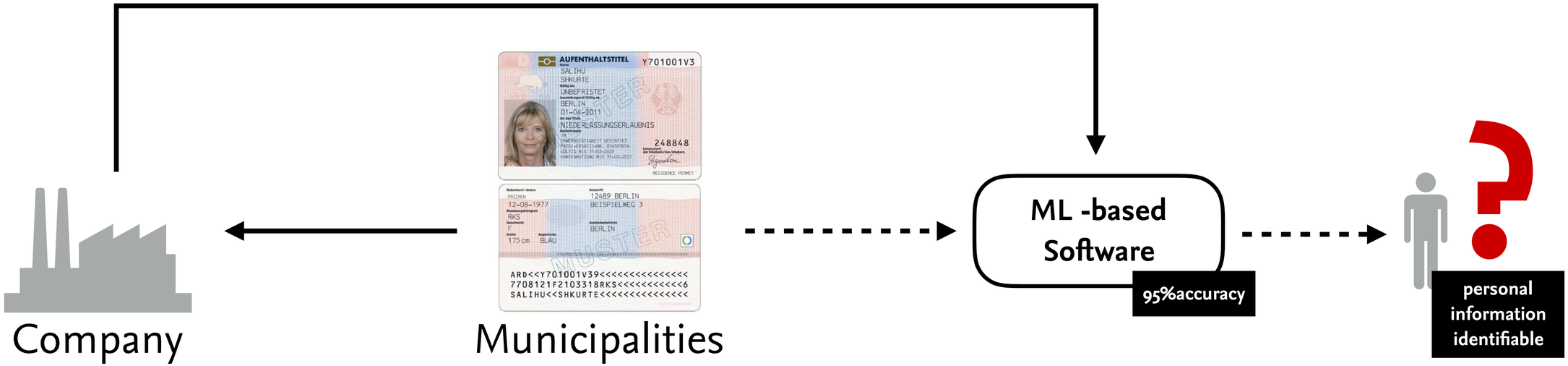
Principles

- Professional Responsibility
- Promotion of Human Values
- Fairness and Non-discrimination
- Human Control of Technology
- Safety and Security
- Transparency and Explainability
- Accountability
- Privacy



Accountability

Use Case: Automatic Anonymization



> ’Algorithmic Accountability’ is the assessment of algorithms to discern and track bias, presuppositions, and prejudices built into, or resulting from algorithms affecting the society.

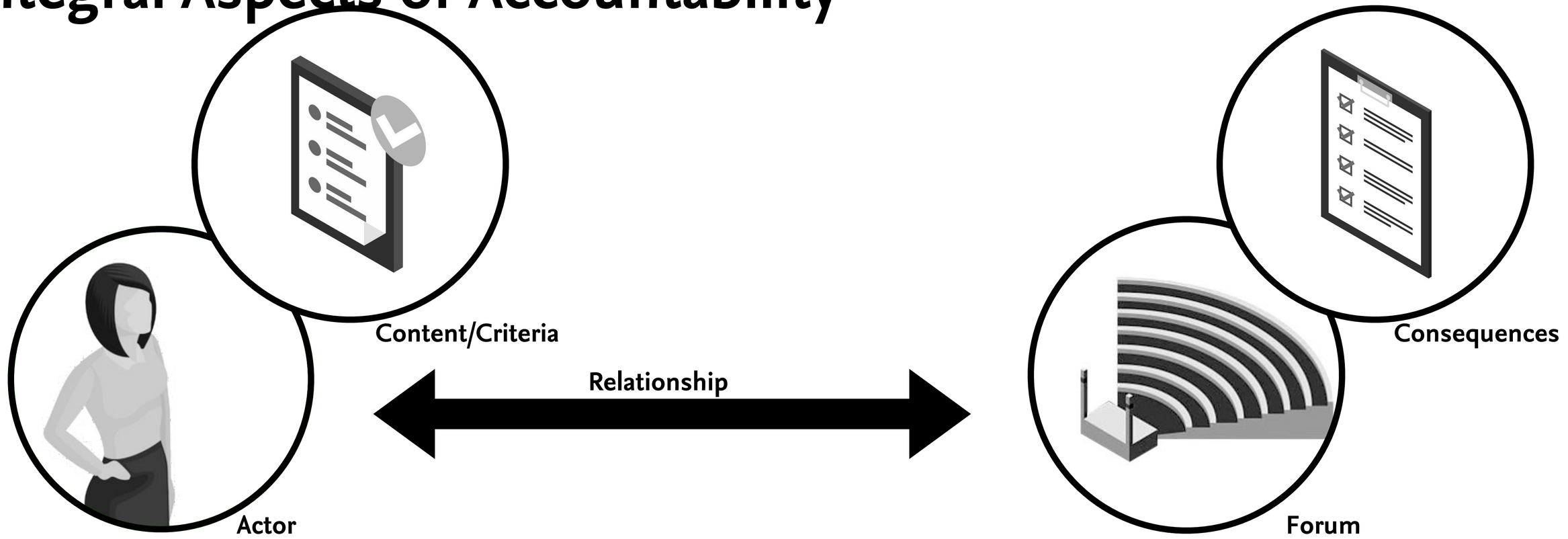
Wieringa, M. (2020). What to account for when accounting for algorithms (Vol. 40, pp. 1–18). Presented at the FAT* '20: Conference on Fairness, Accountability, and Transparency, New York, NY, USA: ACM.

Defining Accountability

> Accountability is: “a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences.”

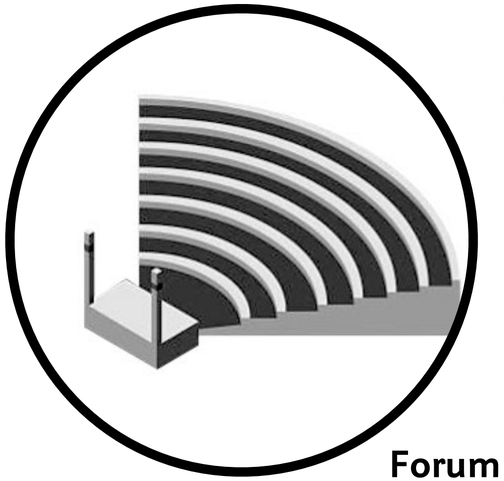
Mark Bovens. (2007) Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal 13, 4 (2007), 447–468.

Integral Aspects of Accountability



Mark Bovens. (2007) Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal 13, 4 (2007), 447–468.

Accountability Relations by Forum



Forum

Political accountability (e.g. ministerial responsibility)

Legal accountability (e.g. judges)

Administrative accountability
(e.g. auditors inspecting a system)

Professional accountability (e.g. insight by peers)

Social accountability (e.g. civil society)

Mark Bovens. (2007) Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal 13, 4 (2007), 447–468.

Photo by Kenny Eliason on Unsplash

MIT Technology Review

Topics+ The Download

Connectivity

The Growing Problem of Bots That Fight Online

The way software agents interact on the Web is poorly understood. Now evidence shows that they fight each other for years.

by Emerging Technology from the arXiv September 20, 2016

become a supporter | subscribe / find a job

theguardian

news / opinion / sport / arts / life

tech / world / UK / science / cities / global development / more

Wikipedia

Study reveals bot-on-bot editing wars raging on Wikipedia's pages

HUFFPOST

INNOVATION

Automated Wikipedia Edit-Bots Have Been Fighting Each Other For A Decade

In 13 different languages.

POPULAR SCIENCE

Technology

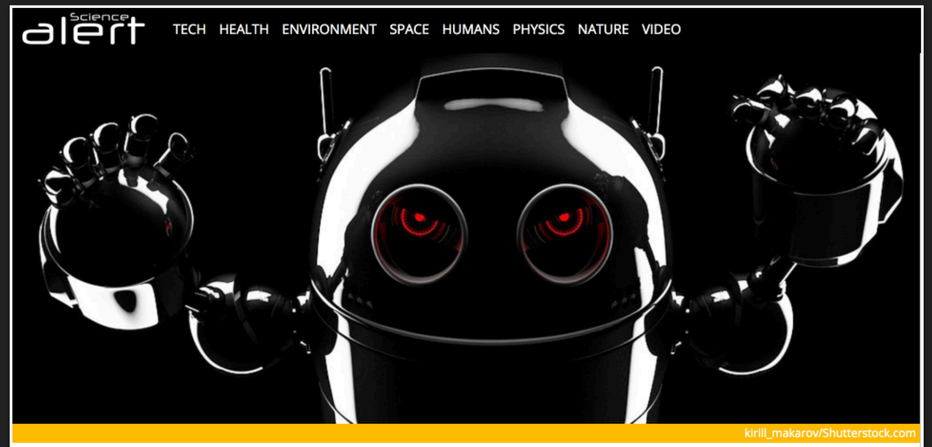
Wikipedia bots spent years fighting silent, tiny battles with each other

And no one even noticed

By Sara Chodosh February 27, 2017

ScienceAlert

TECH HEALTH ENVIRONMENT SPACE HUMANS PHYSICS NATURE VIDEO



kirill_makarov/Shutterstock.com

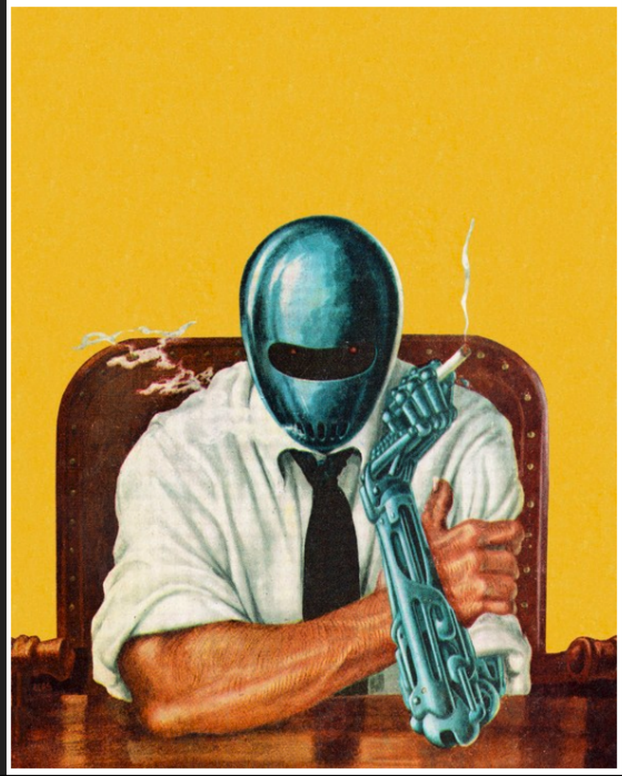
Investigation Reveals That Wikipedia's Bots Are in a Silent, Never-Ending War With Each Other

Technology can be brutal.

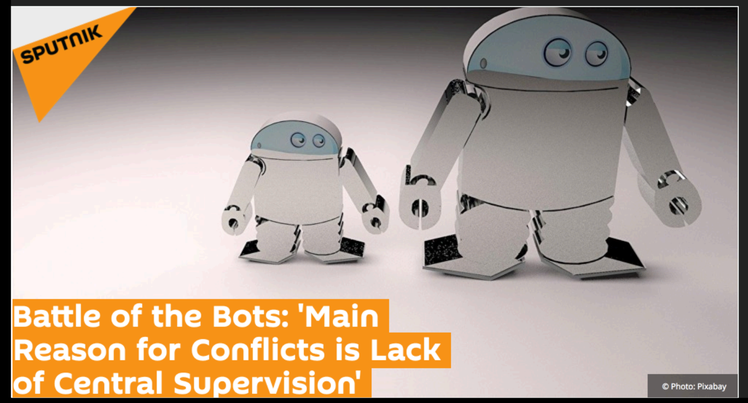
PETER DOCKRILL 26 FEB 2017

MATT SIMON BUSINESS 03.01.17 07:00 AM

INTERNET BOTS FIGHT EACH OTHER BECAUSE THEY'RE ALL TOO HUMAN



SPUTNIK



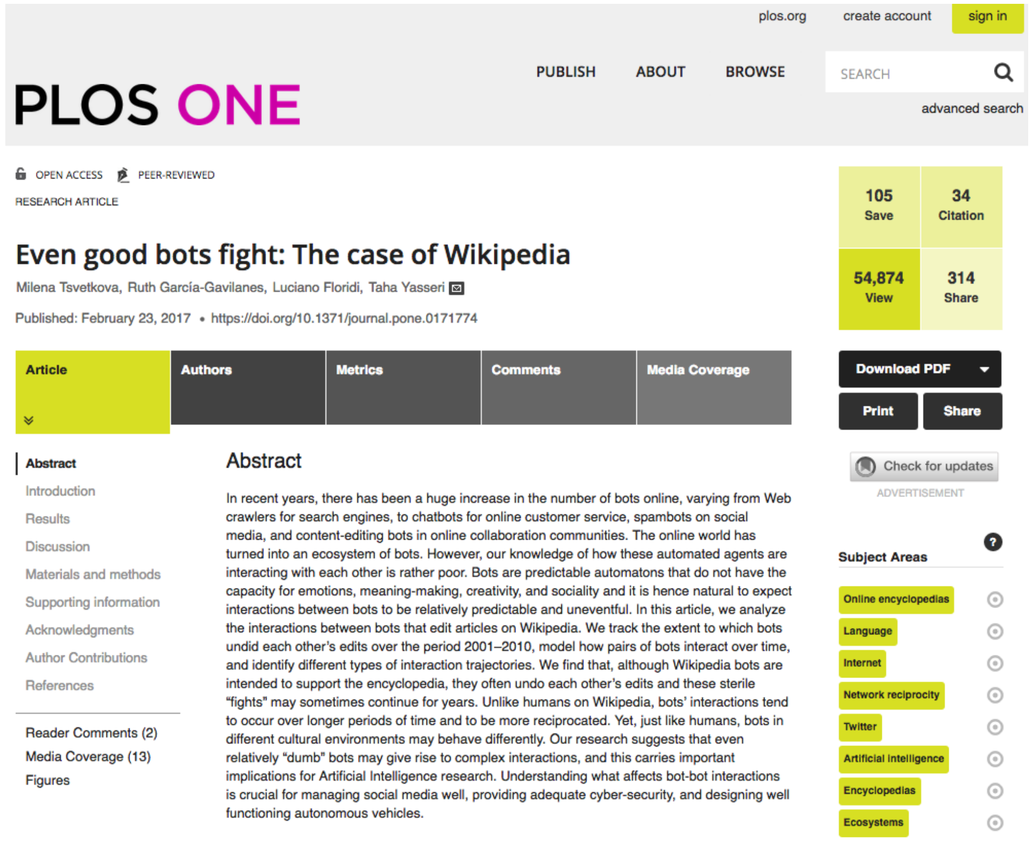
Battle of the Bots: 'Main Reason for Conflicts is Lack of Central Supervision'

© Photo: Pixabay

Home 〉 'Computer bots are like humans, having fights lasting years'

'Computer bots are like humans, having fights lasting years'

Use Case: Even Good Bots Fight



plos.org create account sign in

PLOS ONE

PUBLISH ABOUT BROWSE SEARCH advanced search

OPEN ACCESS PEER-REVIEWED
RESEARCH ARTICLE

Even good bots fight: The case of Wikipedia

Milena Tsvetkova, Ruth García-Gavilanes, Luciano Floridi, Taha Yasseri

Published: February 23, 2017 • https://doi.org/10.1371/journal.pone.0171774

105 Save | 34 Citation | 54,874 View | 314 Share

Article | Authors | Metrics | Comments | Media Coverage

Download PDF | Print | Share

Abstract
Introduction
Results
Discussion
Materials and methods
Supporting information
Acknowledgments
Author Contributions
References

Reader Comments (2)
Media Coverage (13)
Figures

Abstract

In recent years, there has been a huge increase in the number of bots online, varying from Web crawlers for search engines, to chatbots for online customer service, spambots on social media, and content-editing bots in online collaboration communities. The online world has turned into an ecosystem of bots. However, our knowledge of how these automated agents are interacting with each other is rather poor. Bots are predictable automatons that do not have the capacity for emotions, meaning-making, creativity, and sociality and it is hence natural to expect interactions between bots to be relatively predictable and uneventful. In this article, we analyze the interactions between bots that edit articles on Wikipedia. We track the extent to which bots undid each other's edits over the period 2001–2010, model how pairs of bots interact over time, and identify different types of interaction trajectories. We find that, although Wikipedia bots are intended to support the encyclopedia, they often undo each other's edits and these sterile "fights" may sometimes continue for years. Unlike humans on Wikipedia, bots' interactions tend to occur over longer periods of time and to be more reciprocated. Yet, just like humans, bots in different cultural environments may behave differently. Our research suggests that even relatively "dumb" bots may give rise to complex interactions, and this carries important implications for Artificial Intelligence research. Understanding what affects bot-bot interactions is crucial for managing social media well, providing adequate cyber-security, and designing well functioning autonomous vehicles.

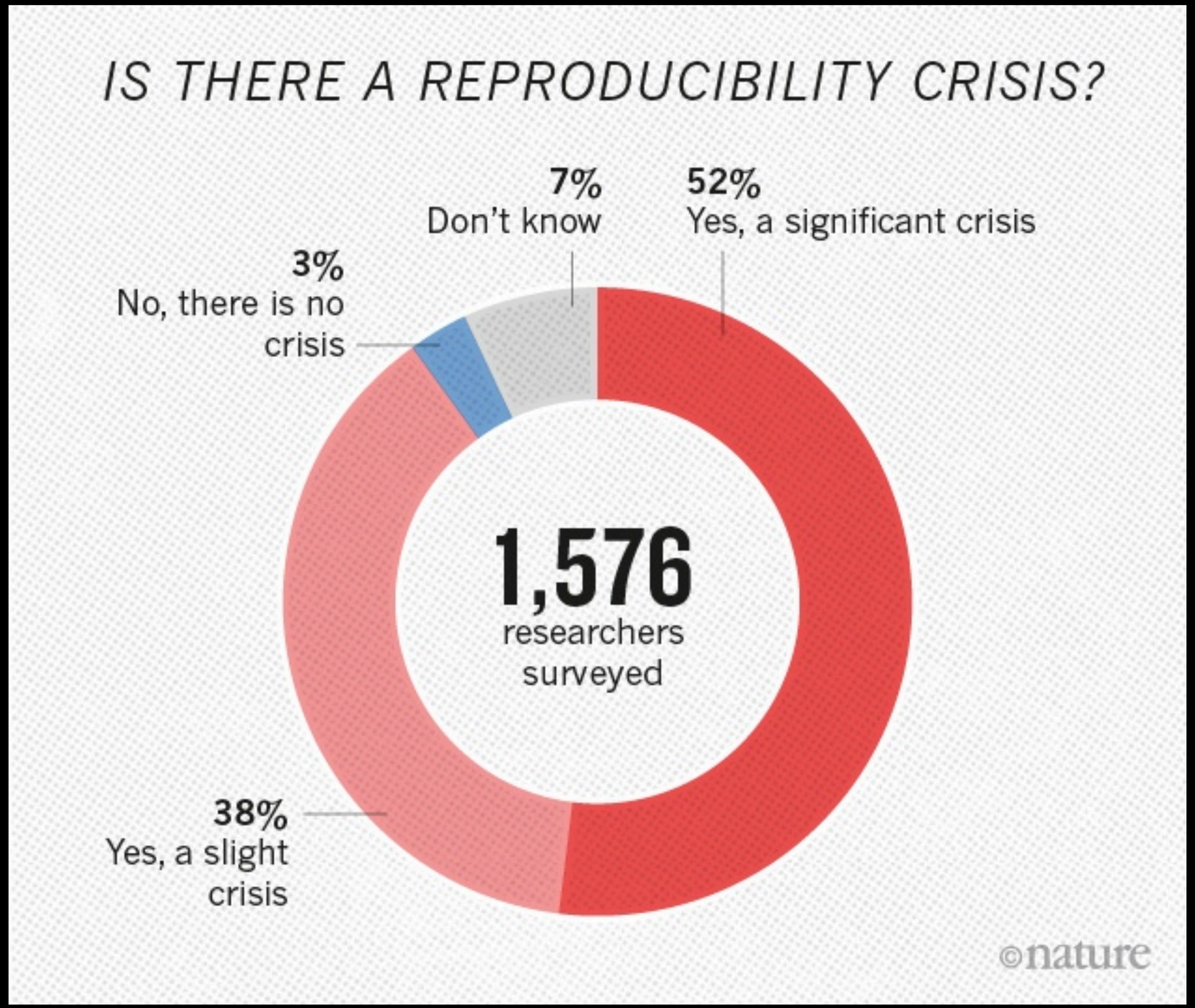
Check for updates
ADVERTISEMENT

Subject Areas
Online encyclopedias
Language
Internet
Network reciprocity
Twitter
Artificial intelligence
Encyclopedias
Ecosystems

Abstract

In recent years, there has been a huge increase in the number of bots online […]. The online world has turned into an ecosystem of bots. However, our knowledge of how these automated agents are interacting with each other is rather poor. […] In this article, we analyze the interactions between bots that edit articles on Wikipedia. We track the extent to which bots undid each other's edits over the period 2001–2010, model how pairs of bots interact over time, and identify different types of interaction trajectories. We find that, although Wikipedia bots are intended to support the encyclopedia, they often undo each other's edits and these sterile "fights" may sometimes continue for years.[…]

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of Wikipedia. PLoS ONE, 12(2)



Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature News, 533(7604), 452.

Replication Study by Geiger & Halfaker

ACM DL DIGITAL LIBRARY

R. Stuart Geiger and Aaron Halfaker. 2017. Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 49 (November 2017), 33 pages. DOI:https://doi.org/10.1145/3134684

What is Reproducibility?

> A research project is computationally reproducible if a second investigator (including you in the future) can recreate the final reported results of the project, including key quantitative findings, tables, and figures, given only a set of files and written instructions.

Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press. http://www.practicereproducibleresearch.org/

Repeatability vs. Reproducibility vs. Replicability

Repeatability — same team, same experimental setup

»[…] For computational experiments, this means that a researcher can reliably repeat her own computation.

Reproducibility — different team, same experimental setup

»[…] For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.

Replicability — different team, different experimental setup

»[…] For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

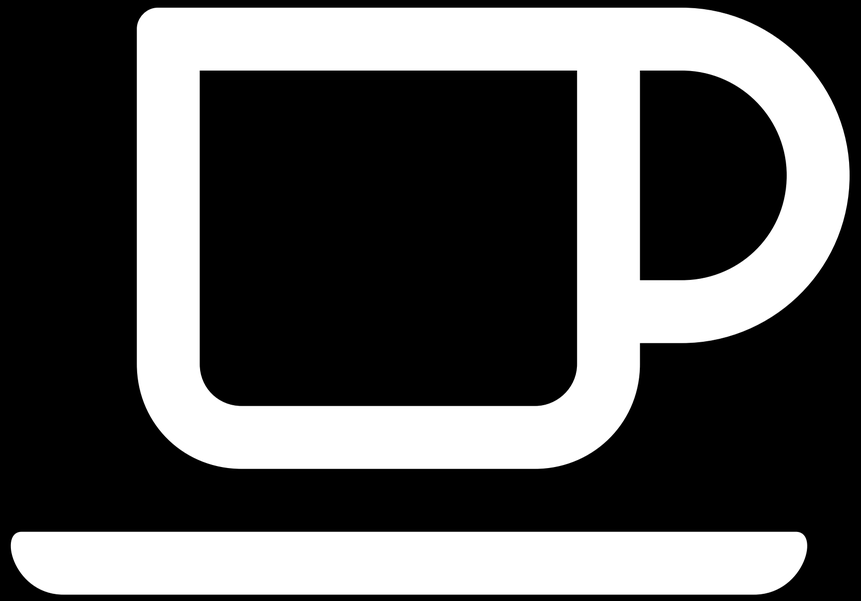
https://www.acm.org/publications/policies/artifact-review-and-badging-current

Degree of Replication

	Same Measurement and Analysis	Different Measurement and Analysis
Same Data Set	1. Checking the analysis	2. Re-analyzing the data
Same Population	3. Exact replication	4. Conceptual extension
Different Population	5. Empirical generalization	6. Generalization and extension

Park, C. L. (2004). What is the value of replicating other studies?. Research Evaluation, 13(3), 189-195.

Translating this Principle in your Data Science Practice



5 minutes break

Translating this Principle in your Data Science Practice

Why is it useful to think about Reproducibility?

Publishing your research openly shows that you take responsibility for your research. Including the possibility that you might be wrong.

It provides transparency around your values, motivations, and assumptions.

Having a public audience in mind when designing and publishing your projects encourages you to reflect on your values, motivations, assumptions, and thought process—and how that might influence your project.

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

Taking a Human-Centered Approach

Values →

Audience Who are you publishing your research for?

Purpose How do you want them to use your research?

Context What factors (under your control) will impact whether/how they use it?

Thinking in terms of "human-centeredness" can help you think of trade-offs of opening your data. For example, it can help you decide when/what NOT to publish openly!

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

What should you Consider in Your Data Science Practice?

1. **Support using your data**

   Make your collected data freely available by considering the principles of FAIR data.

2. **Support sharing your research**

   License your data/scripts/software.

3. **Support understanding your research**

   Prepare your methods and the entire process of data analysis as reproducible as possible and provide relevant documentation.

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

Principles of FAIR Data

Findable

Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.

Accessible

Once the user finds the required data, she/he needs to know how they can be accessed, possibly including authentication and authorization.

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data.

Upload Your Data with a DOI

Making your project identifiable via a Digital Object Identifiers (DOIs)

- Unique ID for an artefact
- Works even if sites fall over
- Single point of reference

Comment: Why uploading data, when people can ask you?

100% of the authors in the following studies signed to share the data upon request.

Actual sharing rate

- »**27%** out of 141 requests (Wicherts et al., 2006)
- »**38%** out of 394 requests (Vanpaemel et al., 2015)
- »**44%** out of 204 requests provided some „artifacts“ (Stodden et al., 2018), where 26% could be reproduced.

„Sharing upon request“ does not work.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). http://doi.org/10.1037/0003-066X.61.7.726 | Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). http://doi.org/10.1525/collabra.13
Stodden, V., Seiler, J., & Ma, Z. (2018). http://doi.org/10.1073/pnas.1708290115 | Slide adapted from Felix Schönbrodt “Changing incentive structures to foster the actual sharing rate of open data” (2018)

What should you consider in your data science practice?

1. **Support using your data**

   Make your collected data freely available by considering the principles of FAIR data.

2. **Support sharing your research**

   License your data/scripts/software.

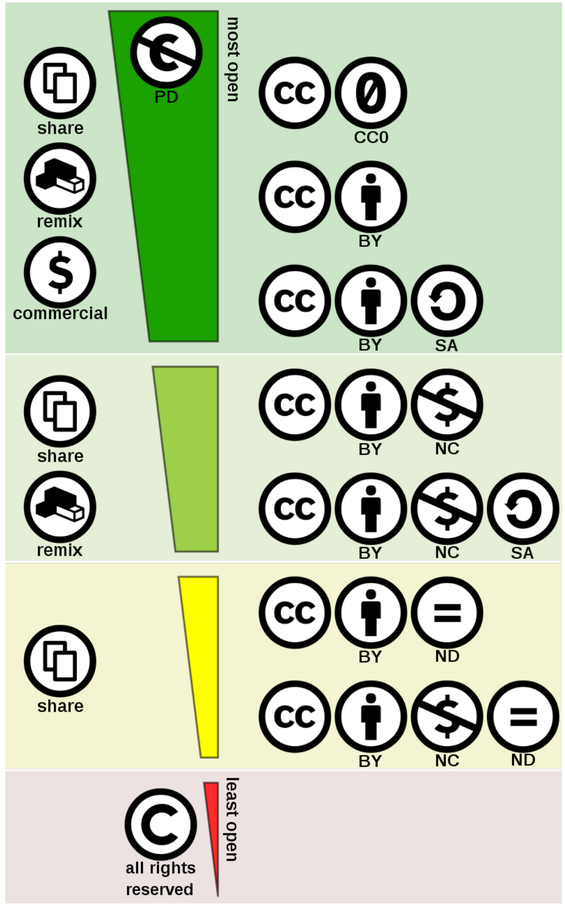
3. **Support understanding your research**

   Prepare your methods and the entire process of data analysis as reproducible as possible and provide relevant documentation.

License Your Data and Documentation

For **licensing data or documentation**, one of the Creative Commons (CC) licenses can be employed.

It allows the creator of the data/documentation, to give other people the right to share, use, and build upon the data/documentation.



Common building blocks:

- **0:** public domain release
- **BY:** must provide attribution
- **NC:** cannot use commercially
- **ND:** cannot make derivatives
- **SA:** must release derivatives under the same/a compatible license

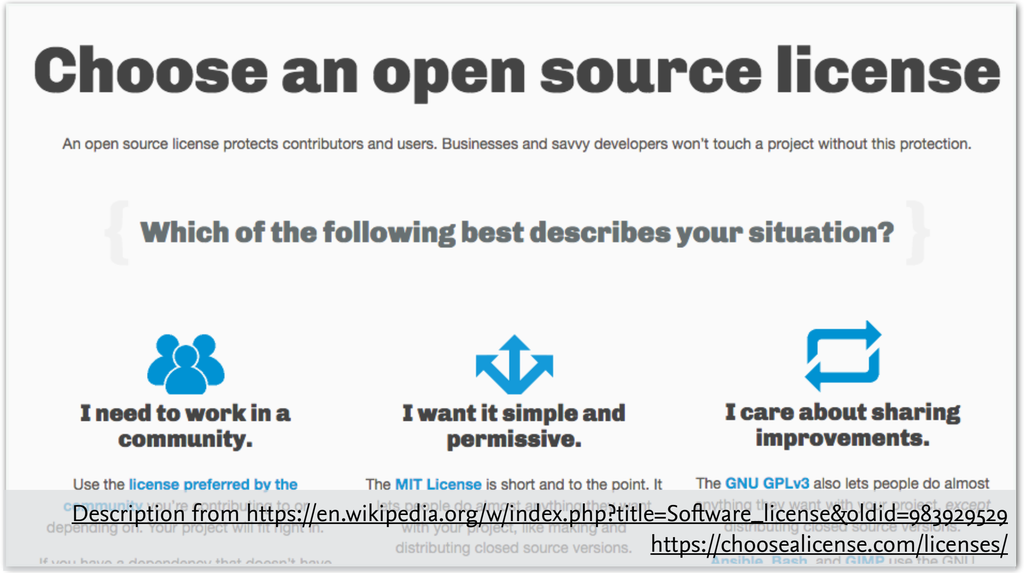
Image by Shaddim, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=47247325
Description from https://en.wikipedia.org/w/index.php?title=Creative_Commons_license&oldid=984371920

License Your Software

A **software is a legal instrument** governing the use or redistribution of software.

There are **many different open source licenses** available such as the MIT license or the General Public License v3.0. But there are quite different, while the MIT License lets people do almost anything they want with your source code, the GPL v3.0 is much stricter.

As a rule of thumb, If you extend existing code, use the license of the community in which you are participating.



Description from https://en.wikipedia.org/w/index.php?title=Software_license&oldid=983929529
https://choosealicense.com/licenses/

Example

Instructional Material

All Software Carpentry and Data Carpentry instructional material is made available under the Creative Commons Attribution license. The following is a human-readable summary of (and not a substitute for) the full legal text of the CC BY 4.0 license.

You are free:

- to **Share**—copy and redistribute the material in any medium or format
- to **Adapt**—remix, transform, and build upon the material

for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

- **Attribution**—You must give appropriate credit (mentioning that your work is derived from work that is Copyright © Software Carpentry and, where practical, linking to http://software-carpentry.org/), provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions—You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits. With the understanding that:

Notices:

- You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.
- No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

Software

Except where otherwise noted, the example programs and other software provided by Software Carpentry and Data Carpentry are made available under the OSI-approved MIT license.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Trademark

"Software Carpentry" and "Data Carpentry" and their respective logos are registered trademarks of NumFOCUS.

https://reproducible-science-curriculum.github.io/sharing-RR-Jupyter/LICENSE.html

Copyright © 2016–2019 Data Carpentry

Edit on GitHub / Contributing / Source / Cite / Contact

What should you consider in your data science practice?

1. **Support using your data**

   Make your collected data freely available by considering the principles of FAIR data.

2. **Support sharing your research**

   License your data/scripts/software.

3. **Support understanding your research**

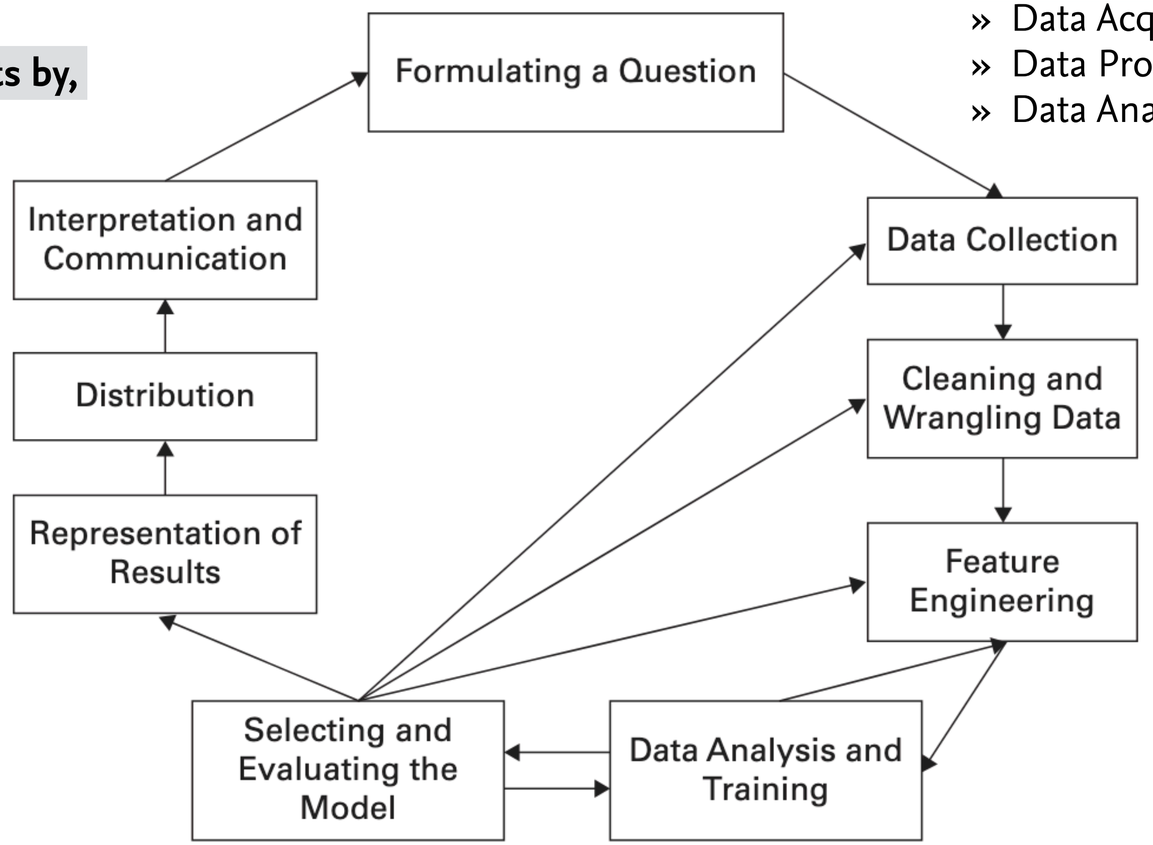
   Prepare your methods and the entire process of data analysis as reproducible as possible and provide relevant documentation.

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

Stages of a Basic Reproducible Workflow

Re-consider the stages of a reproducible workflow
» Data Acquisition
» Data Processing
» Data Analysis

Analyze and derive improvements by, e.g.
» Bug tracker
» Interviews
» Checklists

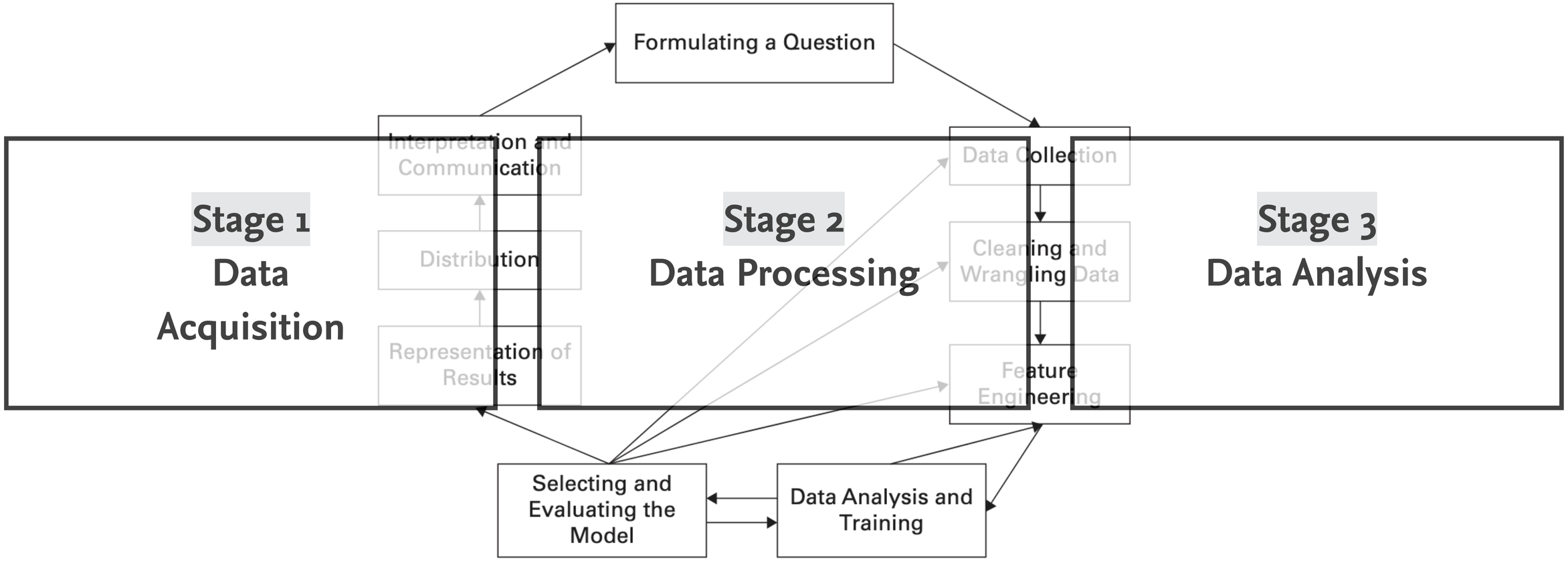


Build new/Refactor existing workflow by, e.g.
» Defining data origin and license
» Documenting used libraries
» Providing test data

Open workflow for feedback of, e.g.
» Peers
» Colleagues
» Community

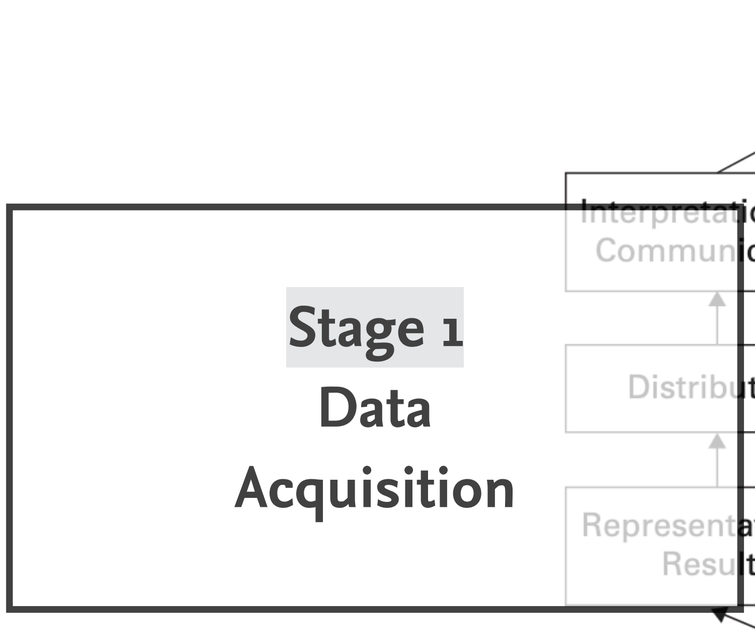
Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press.

Stages of the Basic Reproducible Workflow



Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press.

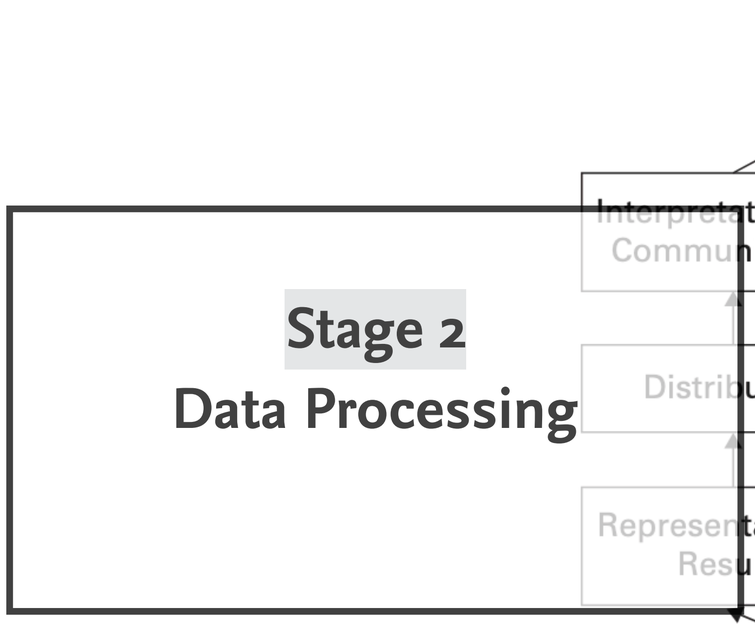
Stages of the Basic Reproducible Workflow



- ✓ Where is your data coming from?
- ✓ What licenses apply to the collected data?
- ✓ Who collected your data?
- ✓ What client-side tools/environments were used to collect your data?
- ✓ If your data is a sample, what were the parameters?
- ✓ What features are described in your data?
- ✓ Is a local copy of your source data available?
- ✓ Are there known errors, inconsistencies, or incompleteness in your source data?

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

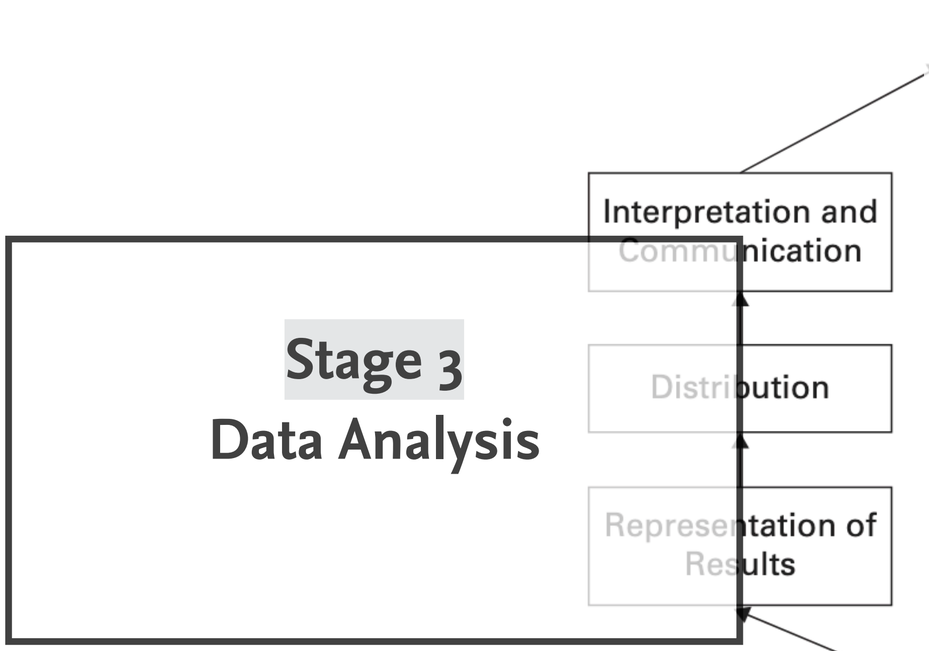
Stages of the Basic Reproducible Workflow



- ✓ What tools/libraries/environments were used in the processing of your data?
- ✓ What sub-sampling, filtering, aggregation, or transformation steps were performed?
- ✓ What order were processing steps performed in?
- ✓ Why were these processing steps performed?
- ✓ How were errors, inconsistencies, or incompletes discovered, and how were they addressed?
- ✓ Did your data processing involve any manual (i.e. non-programmatic) steps?
- ✓ Are you making incremental datasets available?
- ✓ Are you making a final processed dataset available?

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

Stages of the Basic Reproducible Workflow



- ✓ What are the goals of your analysis?
- ✓ What is the nature of your analysis?
- ✓ What assumptions about your data are required for your analysis?
- ✓ What tools/libraries/environments were used in the analysis of your data?
- ✓ What order were analysis steps performed in?
- ✓ Why and how was each analytical step performed?
- ✓ Are you making samples, demos, or test sets available?
- ✓ How are the results of your analysis presented?
- ✓ Are you making a final analyzed dataset available?

Slide adapted from Jonathan T. Morgan & Oliver Keyes (Course DATA 512)

Three Key Practices for your Reproducible Workflow

1. Clearly separate, label, and document all data, files, and operations that occur on data and files.

2. Document all operations fully, automating them as much as possible, and avoiding manual intervention in the workflow when feasible.

3. Design a workflow as a sequence of small steps that are glued together, with intermediate outputs from one step feeding into the next step as inputs.

Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences. Oakland, CA: University of California Press.

Use Case (continued)

Article and Data provided by Tsvetkova et al.

Data Availability: All relevant data are available from figshare at 10.6084/m9.figshare.4597918.

PLOS ONE

Attribution 4.0 International (CC BY 4.0)

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of Wikipedia. PLoS ONE, 12(2).

Data provided by Tsvetkova et al. article

de_wiki.txt — all

```
1   Berserk
2   ^^^_2011-08-18T09:11:59Z 0 261 Don-kun
3   ^^^_2011-08-03T18:45:02Z 0 260 Aka
4   ^^^_2011-07-27T21:19:09Z 0 259 89.166.166.106
5   ^^^_2011-07-27T21:14:45Z 0 258 89.166.166.106
6   ^^^_2011-07-08T07:31:43Z 0 257 Don-kun
7   ^^^_2011-07-08T07:26:32Z 0 256 84.134.22.64
8   ^^^_2011-07-08T07:26:05Z 0 255 84.134.22.64
9   ^^^_2011-05-02T09:15:37Z 0 254 Onkel74
10  ^^^_2011-04-10T11:31:10Z 0 253 Timk70
11  ^^^_2011-04-10T11:26:28Z 0 252 Timk70
12  ^^^_2011-04-06T16:48:26Z 0 251 Marcus_Cyron
13  ^^^_2011-04-06T16:37:22Z 0 250 Don-kun
14  ^^^_2011-04-06T16:01:06Z 0 249 84.178.9.50
15  ^^^_2011-03-04T09:34:15Z 0 248 Gary_Dee
16  ^^^_2011-02-28T01:44:39Z 0 247 Luckas-bot
17  ^^^_2011-02-22T19:42:41Z 0 246 80.187.102.46
18  ^^^_2011-02-22T03:17:41Z 0 245 Mohlemueller
19  ^^^_2011-02-17T14:56:27Z 0 244 Don-kun
20  ^^^_2010-12-17T21:15:08Z 0 243 188.195.22.70
21  ^^^_2010-11-03T15:35:54Z 0 242 LucienBOT
22  ^^^_2010-10-24T21:25:29Z 1 240 83.189.43.244
23  ^^^_2010-10-24T21:25:01Z 0 241 83.189.43.244
24  ^^^_2010-08-08T18:11:04Z 0 240 Mps
25  ^^^_2010-06-22T11:39:40Z 0 239 VolkovBot
26  ^^^_2010-06-15T06:17:54Z 0 238 Xqbot
27  ^^^_2010-04-27T22:17:40Z 1 236 Marcus_Cyron
28  ^^^_2010-04-27T17:33:18Z 0 237 93.83.18.218
29  ^^^_2010-04-12T16:08:16Z 0 236 Niabot
30  ^^^_2010-04-12T16:03:45Z 0 235 Don-kun
```

Line: 1 Plain Text Tab Size: 4

```
^^^_2010-11-03T15:35:54Z 0 242 LucienBOT
^^^_2010-10-24T21:25:29Z 1 240 83.189.43.244
^^^_2010-10-24T21:25:01Z 0 241 83.189.43.244
^^^_2010-08-08T18:11:04Z 0 240 Mps
^^^_2010-06-22T11:39:40Z 0 239 VolkovBot
^^^_2010-06-15T06:17:54Z 0 238 Xqbot
^^^_2010-04-27T22:17:40Z 1 236 Marcus_Cyron
^^^_2010-04-27T17:33:18Z 0 237 93.83.18.218
^^^_2010-04-12T16:08:16Z 0 236 Niabot
^^^_2010-04-12T16:03:45Z 0 235 Don-kun
```

The data processing step is already incompletely documented. Missing information are the user groups (user, bots). Are vandals excluded?

Replication Study by Geiger & Halfaker

RESEARCH-ARTICLE OPEN ACCESS

Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replicatio and Expansion of 'Even Good Bots Fight'

ACM DL DIGITAL LIBRARY

Authors: R. Stuart Geiger, Aaron Halfaker Authors Info & Affiliations

Publication: Proceedings of the ACM on Human-Computer Interaction • December 2017 • Article No.: 49
• https://doi.org/10.1145/3134684

5 424 eReader PDF

figshare

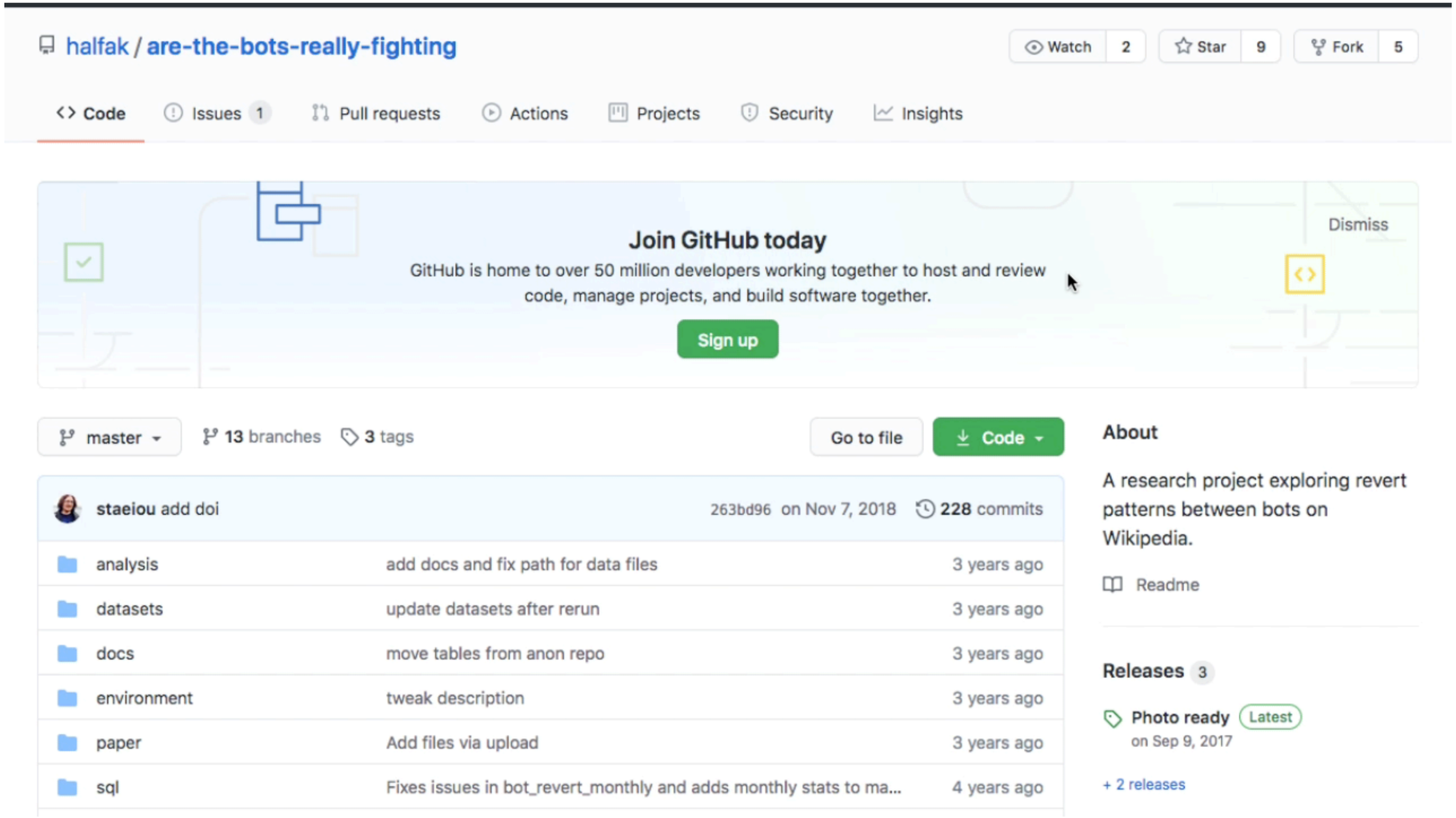
LICENCE

CC0

Code, data, and Jupyter notebooks available at: https://github.com/halfak/are-the-bots-really-fighting
https://doi.org/10.6084/m9.figshare.5362216

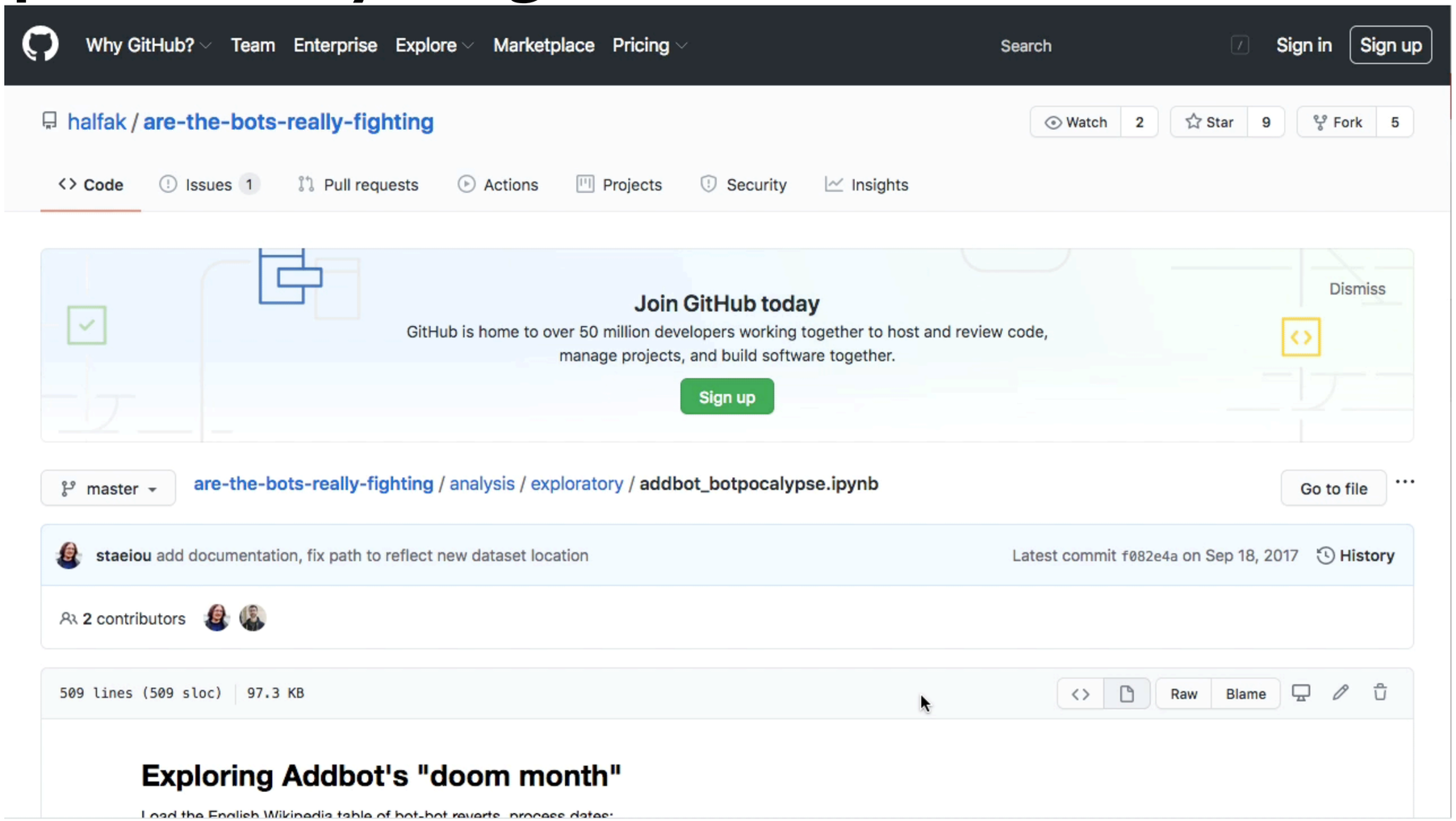
R. Stuart Geiger and Aaron Halfaker. 2017. Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 49 (November 2017), 33 pages. DOI:https://doi.org/10.1145/3134684

Data provided by Geiger & Halfaker



https://github.com/halfak/are-the-bots-really-fighting/

Scripts provided by Geiger & Halfaker



https://github.com/halfak/are-the-bots-really-fighting/

What are the results of Geiger & Halfaker's replication?

They could show that the overwhelming majority of bot-bot reverts constitute routine, productive, and even collaborative work between bots.

They define bots as assemblages of "code and a human developer" which are responsible for operating the bot in alignment with Wikipedia's complex policy environment.

Why these researchers arrived at this very different perspective, we will discuss in another lecture.

Check your Insights

» How do Accountability and Reproducibility relate?

» What is the difference between repeatability, reproducibility, replicability?

» How can you differentiate various forms of replication studies?

» What aspects should you consider in your (open) data science practice?

» What does it mean - FAIR data?

» What is Creative Commons (CC)?

» What should be your key practice of a basic reproducible workflow?

«Human-Centered Data Science»

Next week: Examining the Role of Data in Your Data Science Practice

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

May 19, 2022