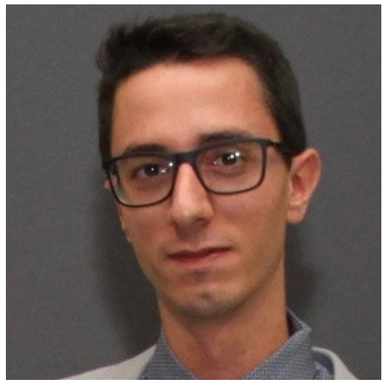
ICWE 2019
19th INTERNATIONAL CONFERENCE ON WEB ENGINEERING
Daejeon, Korea 19th

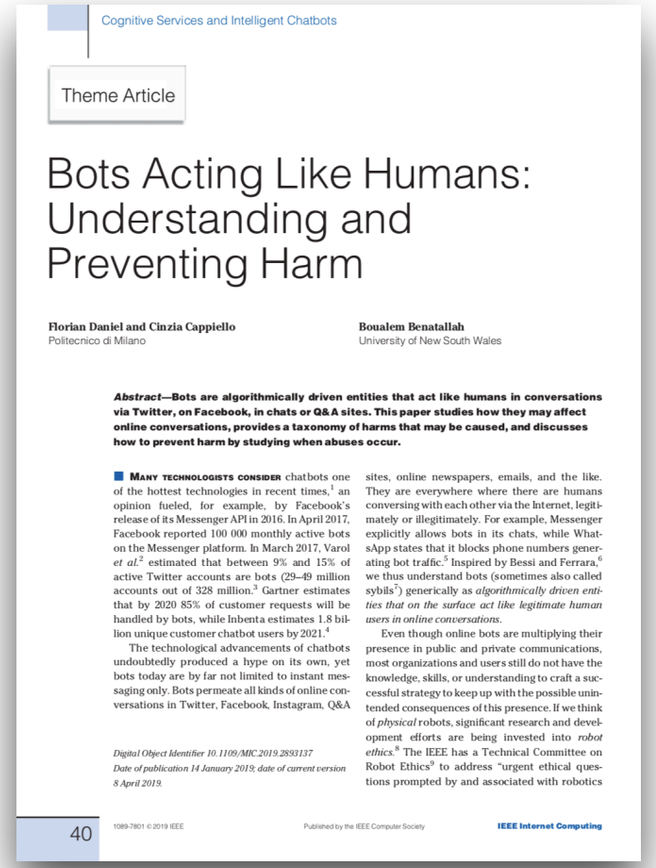
# On Twitter Bots Behaving Badly: Empirical Study of Code Patterns on GitHub


POLITECNICO MILANO 1863
DIPARTIMENTO DI ELETTRONICA INFORMAZIONE E BIOINGEGNERIA



Andrea Millimaggi

Florian Daniel



IEEE Internet Computing

Cognitive Services and Intelligent Chatbots

Theme Article

Bots Acting Like Humans: Understanding and Preventing Harm

Florian Daniel and Cinzia Cappiello
Politecnico di Milano

Boualem Benatallah
University of New South Wales

***Abstract*—Bots are algorithmically driven entities that act like humans in conversations via Twitter, on Facebook, in chats or Q&A sites. This paper studies how they may affect online conversations, provides a taxonomy of harms that may be caused, and discusses how to prevent harm by studying when abuses occur.**

**Many technologists consider** chatbots one of the hottest technologies in recent times,[1] an opinion fueled, for example, by Facebook's release of its Messenger API in 2016. In April 2017, Facebook reported 100 000 monthly active bots on the Messenger platform. In March 2017, Varol *et al.*[2] estimated that between 9% and 15% of active Twitter accounts are bots (29–49 million accounts out of 328 million.[3] Gartner estimates that by 2020 85% of customer requests will be handled by bots, while Inbenta estimates 1.8 billion unique customer chatbot users by 2021.[4]

The technological advancements of chatbots undoubtedly produced a hype on its own, yet bots today are by far not limited to instant messaging only. Bots permeate all kinds of online conversations in Twitter, Facebook, Instagram, Q&A sites, online newspapers, emails, and the like. They are everywhere where there are humans conversing with each other via the Internet, legitimately or illegitimately. For example, Messenger explicitly allows bots in its chats, while WhatsApp states that it blocks phone numbers generating bot traffic.[5] Inspired by Bessi and Ferrara,[6] we thus understand bots (sometimes also called sybils[7]) generically as *algorithmically driven entities that on the surface act like legitimate human users in online conversations*.

Even though online bots are multiplying their presence in public and private communications, most organizations and users still do not have the knowledge, skills, or understanding to craft a successful strategy to keep up with the possible unintended consequences of this presence. If we think of *physical* robots, significant research and development efforts are being invested into *robot ethics*.[8] The IEEE has a Technical Committee on Robot Ethics[9] to address "urgent ethical questions prompted by and associated with robotics

*Digital Object Identifier 10.1109/MIC.2019.2893137*
*Date of publication 14 January 2019; date of current version 8 April 2019.*


1089-7801 © 2019 IEEE
Published by the IEEE Computer Society
IEEE Internet Computing

F. Daniel, C. Cappiello, B. Benatallah. Bots Acting Like Humans: Understanding and Preventing Harm. IEEE Internet Computing 23(2), 2019, Pages 40-49.

# **Harm** in human-bot interactions

## Psychological harm

Someone's psychological health or well-being get endangered or injured

## Legal harm

Someone becomes subject to law enforcement or prosecution

## Economic harm

Someone incurs a monetary cost or loses time

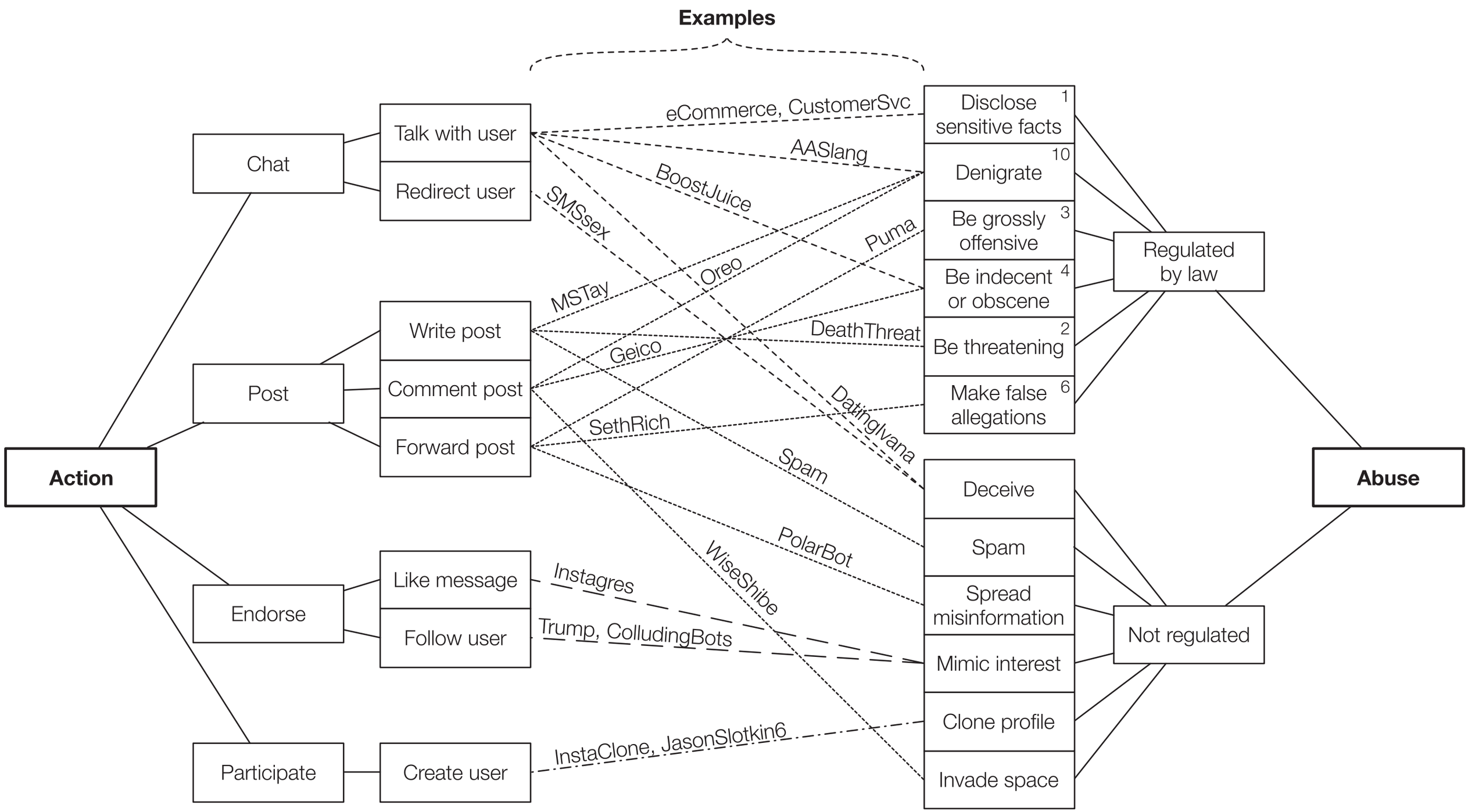
## Social harm

Someone's image or standing in a community gets affected negatively

## Democratic harm

Democratic rules and principles are undermined

# Harm = consequence of an **action** and an **abuse**



**Literature** on bots

Bot development (frameworks, APIs, etc.)

Bot detection from externally visible communications

**Problem** (goal of this paper)

Identify **how** harm is caused by bots

Understand likely underlying **intentions**

>> Abuse-oriented classification of bot **code repositories** published on GitHub

Before going into the details…

Bots are not negative in general!

Platform **policies and permissions**

All platforms provide developers with **programmable interfaces**

Typically allow programmatic access to **all functionalities**

Users of the APIs must **authenticate** with the platforms

Almost all platforms impose some kind of **limitation**

# Dataset

Focus on **Twitter**

Search by keywords

Collection of code files + metadata

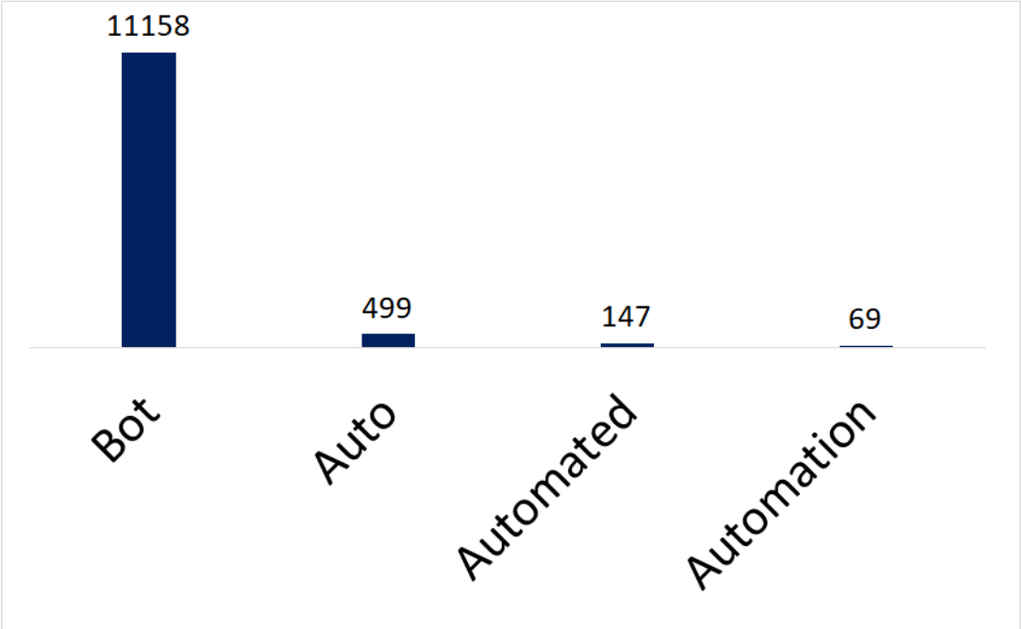



Fig. 1: Distribution of GitHub search results by searched keywords (includes all programming languages).

**Preliminary analysis**
of actions implemented in repositories

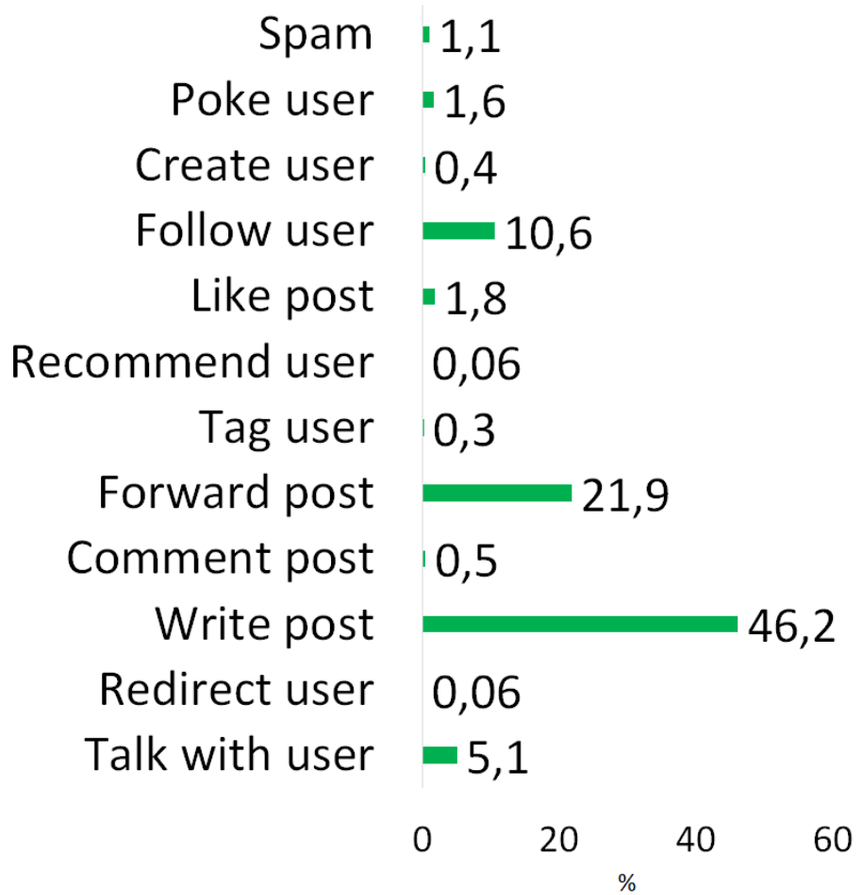



Fig. 2: Labels of repositories.

## Starting point Final dataset

**Selection criteria:**

Programming language = Python

Exclusion of repositories that are out of scope

5 best repositories for each of the most used actions

5 random repositories from the rest

10 best repositories we could not classify

10 random repositories from the rest

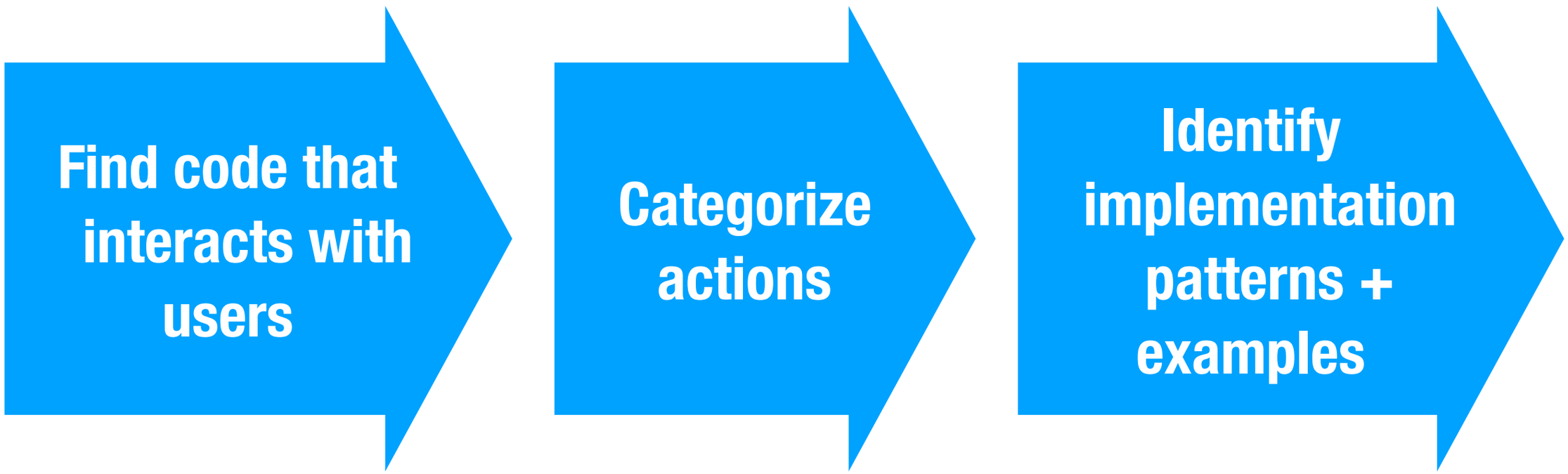
**= 60 GitHub Twitter bot repositories programmed in Python**

average number of files per repository: 3

average number of lines of code: 192

average size of repository: 21.39 KBytes

# Methodology




Systematic, **manual** code review!

# Results **Actions** for participation

| Action | Description |
|---|---|
| *Search* | Search users or tweets using names, keywords, hashtags, ids or similar or by navigating social network relationships (e.g., friends of friends, followers of friends, friends of followers, followers of followers) |
| *Follow* | Follow users to establish social relationships |
| *Like* | Like tweets by other users to endorse them |
| *Tweet* | Post a new tweet to communicate content |
| *Mention* | Mention other users in tweets using @ to attract attention |
| *Retweet* | Re-post tweets by other users to endorse them |
| *Talk to* | Send direct messages to users to converse with them |
| *Pause* | Pause the conversation flow of the bot |
| *Store* | Store content retrieved from the social network for later use |

Synthesis of online communication actions implemented by Twitter bots

# Results **Code patterns** of action implementations

Search

Follow

Like

Tweet

Mention

Retweet

Talk to

Pause

Store

Table 2: Taxonomy of code patterns used for the implementation of actions.

| Action | Pattern | Description |
|---|---|---|
| Search | *User search* | Search user account by name, keyword, id or similar |
| | *Tweet search* | Search tweets by keyword or hashtag |
| | *Trend search* | Search trending topics or hashtags by location |
| Follow | *Indiscriminate follow* | Follow users without checking suitability of users, usernames or content shared |
| | *Whitelist-based follow* | Follow only users whose attributes or tweets match some element of a given whitelist |
| | *Blacklist-based follow* | Don't follow users whose attributes or tweets satisfy one or more criteria specified in a blacklist |
| | *Phantom follow* | Follow users and unfollow them as soon as a given condition is satisfied, e.g, a limit of friends reached or being followed back |
| Like | *Indiscriminate like* | Like tweets without checking suitability of content, user or username |
| | *Whitelist-based like* | Like only tweets by users whose attributes or content match some element of a whitelist |
| | *Blacklist-based like* | Don't like tweets whose attributes or users match an element of a blacklist |
| | *Mass like* | Aggressively like tweets of given users |
| Tweet | *Fixed-content tweet* | The content of the tweet is taken from a fixed, static collection of predefined messages |
| | *AI-generated tweet* | The text of the tweet is automatically generated using AI/NLP tools |
| | *Trusted source tweet* | The content of the tweet is taken from a source that can be considered trusted |
| | *Tweet with opt-in* | Tweets are sent only to people who ask to interact with the bot, sending it a message or mentioning it in a tweet |
| Mention | *Indiscriminate mention* | Mention other users without checking suitability of username or content shared |
| | *Targeted mention* | Classify users on the basis of their tweets and mention them in targeted messages |
| | *Whitelist-based mention* | Mention only users whose attributes match some element of a whitelist |
| | *Blacklist-based mention* | Don't mention users whose attributes match elements of a blacklist |
| Retweet | *Indiscriminate retweet* | Retweet tweets without checking content or username for suitability |
| | *Whitelist-based retweet* | Retweet content only from users whose attributes match some element of a whitelist |
| | *Blacklist-based retweet* | Don't retweet tweets whose attributes or users satisfy some condition expressed in a blacklist |
| | *Mass retweet* | Aggressively retweet multiple tweets by selected users |
| Talk to | *Indiscriminate talk* | Send direct, instant messages to users without checking their suitability |
| | *Talk with opt-in* | Reply only to messages sent to the bot (passive behavior) |
| | *AI-generated talk* | Generate messages using AI/NLP tools |
| | *Fixed-content talk* | Take message from a fixed list of predefined phrases |
| | *Targeted talk* | Classify users based on their tweets or attributes and target message accordingly |
| Pause | *Mimic human* | Use pauses in instant messages to deliver human-like conversation experience to other humans |
| | *Satisfy API constraints* | Use as short as possible pauses just to avoid being blocked by API usage limitations |
| Store | *Store persistently* | Store retrieved content or user information for later use |

| Action | Pattern | Description |
|---|---|---|
| Search | *User search* | Search user account by name, keyword, id or similar |
| | *Tweet search* | Search tweets by keyword or hashtag |
| | *Trend search* | Search trending topics or hashtags by location |
| Follow | *Indiscriminate follow* | Follow users without checking suitability of users, usernames or content shared |
| | *Whitelist-based follow* | Follow only users whose attributes or tweets match some element of a given whitelist |
| | *Blacklist-based follow* | Don't follow users whose attributes or tweets satisfy one or more criteria specified in a blacklist |
| | *Phantom follow* | Follow users and unfollow them as soon as a given condition is satisfied, e.g, a limit of friends reached or being followed back |
| Like | *Indiscriminate like* | Like tweets without checking suitability of content, user or username |
| | *Whitelist-based like* | Like only tweets by users whose attributes or content match some element of a whitelist |
| | *Blacklist-based like* | Don't like tweets whose attributes or users match an element of a blacklist |
| | *Mass like* | Aggressively like tweets of given users |
| Tweet | *Fixed-content tweet* | The content of the tweet is taken from a fixed, static collection of predefined messages |
| | *AI-generated tweet* | The text of the tweet is automatically generated using AI/NLP tools |
| | *Trusted source tweet* | The content of the tweet is taken from a source that can be considered trusted |
| | *Tweet with opt-in* | Tweets are sent only to people who ask to interact with the bot, sending it a message or mentioning it in a tweet |

# Example

Library

Action 1: search

```
for tweet in tweepy.Cursor(api.search, q=QUERY).items():
    tweet.user.follow()
```

Action 2: follow

= search users pattern + indiscriminate follow pattern

Function definition

Mention search

```
def mentions(count, max_seconds_ago, id_blacklist) :
    return [mention for mention in api.mentions_timeline(count=count)
      if not mention.id in id_blacklist ]
```

Blacklist inclusion check

= blacklist-based mention pattern

# Results Effects

Patterns may

**Enable** an abuse

↔ Logic that by design performs an abuse

**Prevent** an abuse

↔ Logic that prevents the bot from performing an abuse

Be vulnerable to **content abuse**

↔ Interactions with users and/or content that may be inappropriate

Be vulnerable to **trust abuse**

↔ Forward, store or analyze content retrieved from users

# Results

## Pattern-effect matrix = potential effects of patterns

| Action | Pattern | Disclose sensitive facts | Denigrate | Be grossly offensive | Be indecent or obscene | Be threatening | Make false allegations | Deceive | Spam | Spread misinformation | Mimic interest | Clone profile | Invade space |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Follow | Indiscriminate follow | — | C | C | C | — | — | — | — | — | GO | — | — |
| | Whitelist-based follow | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| | Blacklist-based follow | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| | Phantom follow | — | C | C | C | — | — | — | — | — | GO | — | — |
| Like | Indiscriminate like | — | C | C | C | — | — | — | — | — | GO | — | — |
| | Whitelist-based like | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| | Blacklist-based like | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| | Mass like | — | C | C | C | — | — | — | GO | — | GO | — | — |
| Tweet | Fixed-content tweet | ✋ | ✋ | ✋ | ✋ | ✋ | ✋ | — | GO | — | — | — | — |
| | AI-generated tweet | C | C | C | C | C | C | — | GO | C | — | GO | — |
| | Trusted source tweet | ✋ | ✋ | ✋ | ✋ | ✋ | ✋ | — | GO | — | — | — | — |
| Mention | Indiscriminate mention | — | GO | GO | GO | — | — | — | GO | — | GO | — | — |
| | Opt-in mention | — | GO | GO | GO | — | — | — | ✋ | — | — | — | — |
| | Targeted mention | — | C | — | — | — | C | — | — | — | — | — | — |
| | Whitelist-based mention | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| | Blacklist-based mention | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| Retweet | Indiscriminate retweet | — | C | C | C | C | C | C | — | C | GO | — | — |
| | Whitelist-based retweet | — | ✋ | ✋ | ✋ | ✋ | ✋ | ✋ | — | ✋ | GO | — | — |
| | Blacklist-based retweet | — | ✋ | ✋ | ✋ | ✋ | ✋ | ✋ | — | ✋ | GO | — | — |
| | Mass retweet | — | C | C | C | C | C | C | GO | C | GO | — | — |
| Talk to | Indiscriminate talk | — | — | — | — | — | — | — | GO | — | GO | — | — |
| | Fixed-content talk | ✋ | ✋ | ✋ | ✋ | ✋ | ✋ | — | GO | — | GO | — | — |
| | AI-generated talk | C | C | C | C | C | C | — | GO | C | GO | GO | — |
| | Talk with opt-in | — | — | — | — | — | — | — | ✋ | — | ✋ | — | — |
| | Targeted talk | — | — | — | — | — | — | — | GO | — | GO | — | — |
| Pause | Mimic human | — | — | — | — | — | — | GO | — | — | — | — | — |
| | Satisfy API contraints | — | — | — | — | — | — | — | GO | — | — | — | — |
| Store | Store persistently | T | — | — | — | — | — | — | — | — | — | T | — |

GO *Enables* ✋ *Prevents* C *Vulnerable to content abuse* T *Vulnerable to trust abuse*

# Zoom into **Follow patterns**

| Action | Pattern | Disclose sensitive facts | Denigrate | Be grossly offensive | Be indecent or obscene | Be threatening | Make false allegations | Deceive | Spam | Spread misinformation | Mimic interest | Clone profile | Invade space |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Follow | Indiscriminate follow | — | C | C | C | — | — | — | — | — | GO | — | — |
| | Whitelist-based follow | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| | Blacklist-based follow | — | ✋ | ✋ | ✋ | — | — | — | — | — | GO | — | — |
| | Phantom follow | — | C | C | C | — | — | — | — | — | GO | — | — |

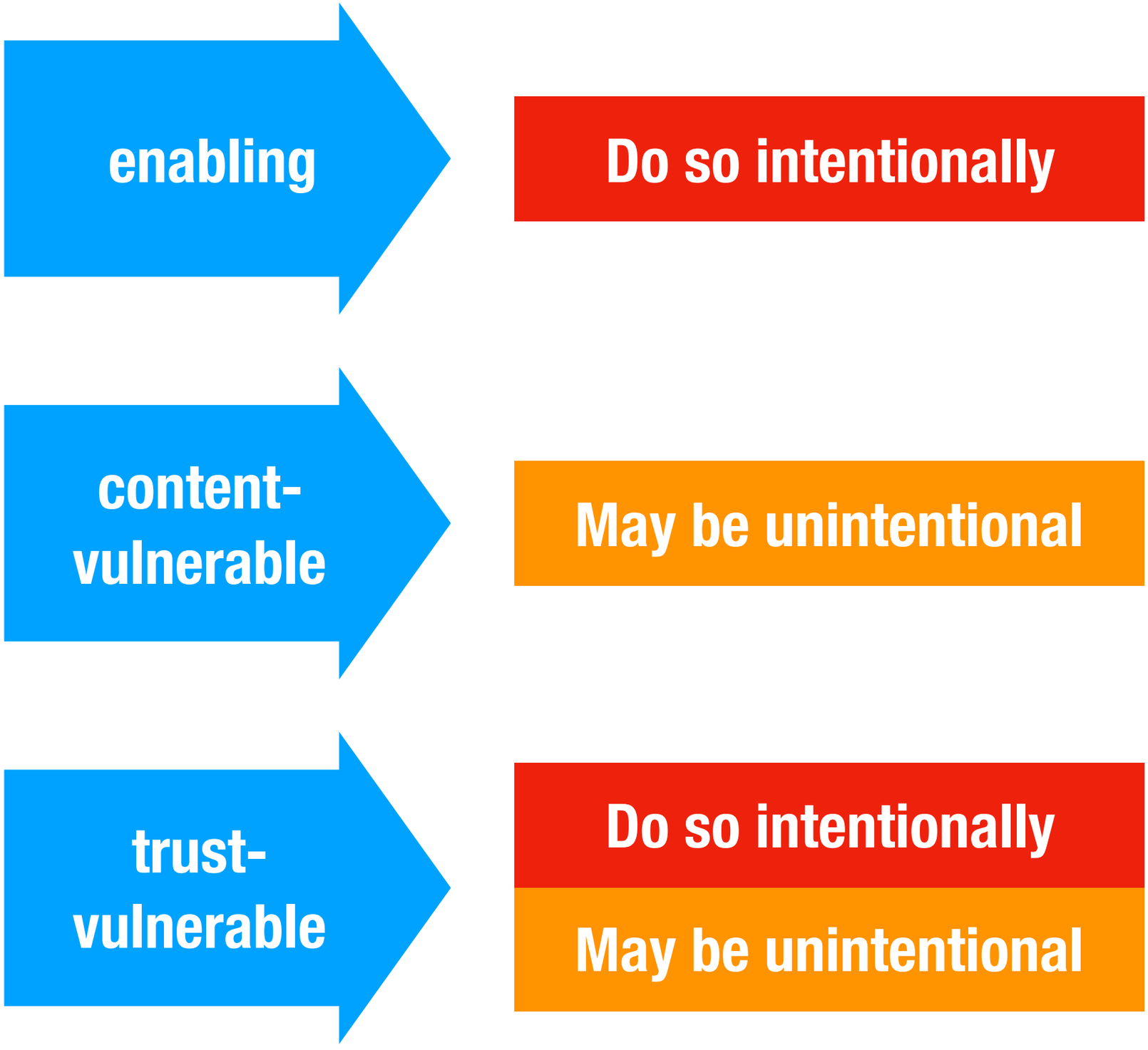
GO *Enables* ✋ *Prevents* C *Vulnerable to content abuse* T *Vulnerable to trust abuse*

# Zoom into **Tweet patterns**

| Action | Pattern | Disclose sensitive facts | Denigrate | Be grossly offensive | Be indecent or obscene | Be threatening | Make false allegations | Deceive | Spam | Spread misinformation | Mimic interest | Clone profile | Invade space |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tweet | Fixed-content tweet | Prevents | Prevents | Prevents | Prevents | Prevents | Prevents | — | GO | — | — | — | — |
| | AI-generated tweet | C | C | C | C | C | C | — | GO | C | — | GO | — |
| | Trusted source tweet | Prevents | Prevents | Prevents | Prevents | Prevents | Prevents | — | GO | — | — | — | — |

GO *Enables* Prevents *Prevents* C *Vulnerable to content abuse* T *Vulnerable to trust abuse*

**Results** Coming back to the "**why**" question… and using some technical considerations on the nature of patterns…



## Summing up

**Original** perspective on bots for online communication: code

Contributions to state of the art:

1. Identified **31 patterns** and 9 actions from 60 repositories (~ 80 hours of manual code review + x of discussion)
2. Discussed **effects** of patterns and mapped patterns to potential abuses
3. Technical interpretation of **intentionality** underlying bot implementations

**Next**: formal language for action patterns + patterns search engine for automated pattern retrieval from all collected repositories