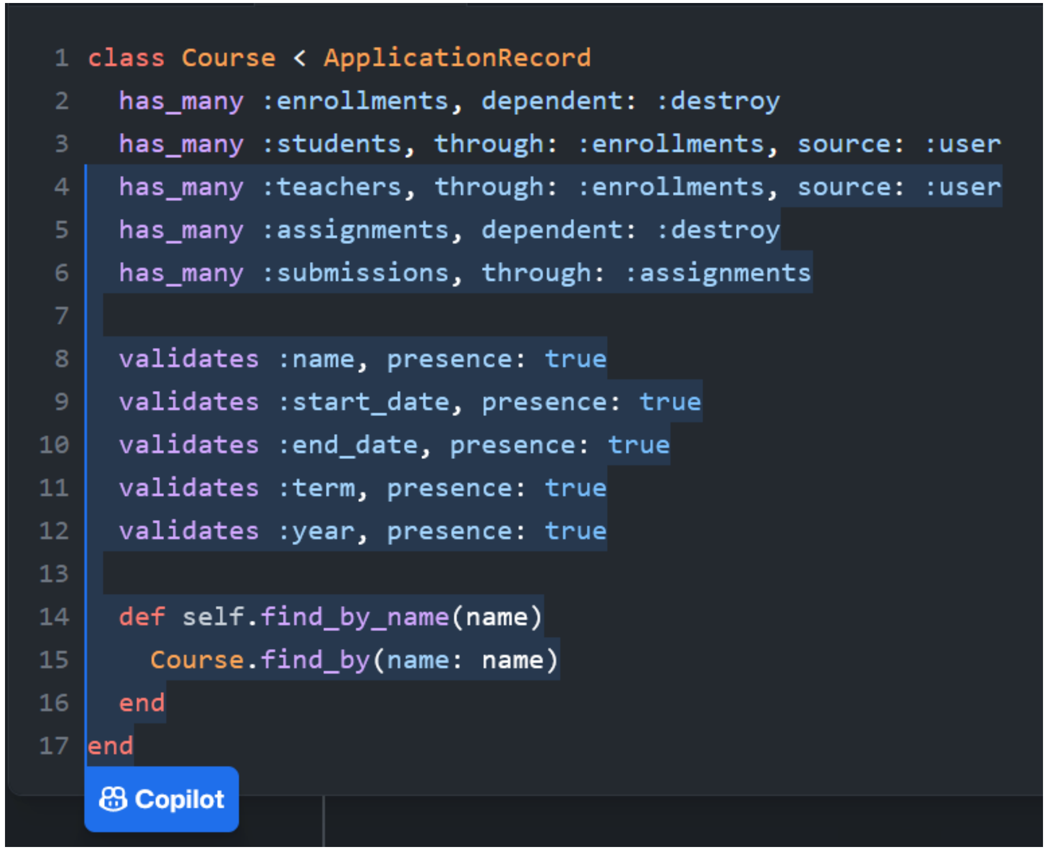
GitHub Copilot a Stable Diffusion – súdne procesy

Martin Tvarožek, 536481

GitHub Copilot

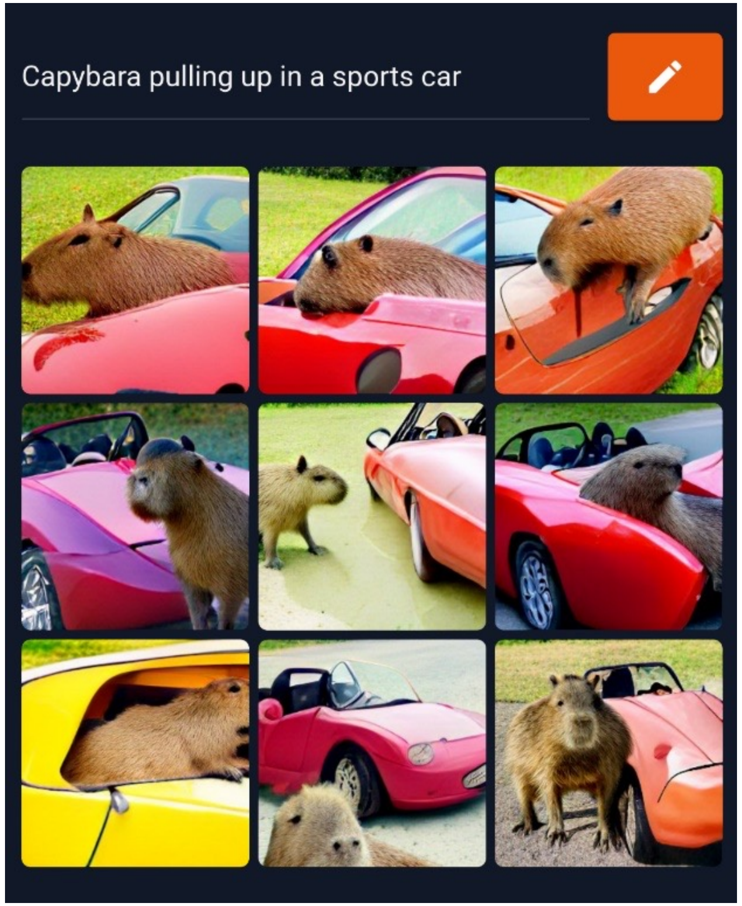
- AI „párový programátor“
- navrhuje alebo automaticky dopĺňa výrazy a funkcie
- trénovaný na verejných GitHub repozitároch a verejne dostupnom kóde
- komerčný produkt ($10/mesiac)



```ruby
class Course < ApplicationRecord
  has_many :enrollments, dependent: :destroy
  has_many :students, through: :enrollments, source: :user
  has_many :teachers, through: :enrollments, source: :user
  has_many :assignments, dependent: :destroy
  has_many :submissions, through: :assignments

  validates :name, presence: true
  validates :start_date, presence: true
  validates :end_date, presence: true
  validates :term, presence: true
  validates :year, presence: true

  def self.find_by_name(name)
    Course.find_by(name: name)
  end
end
```

Stable Diffusion

- Stability AI, Midjourney, Dall-E, DreamUp (DeviantArt)
- skupina text-to-image modelov
- generujú obrázky podľa zadania
- trénované na datasete LAION (Large-scale Artificial Intelligence Open Network)
- mnohé implementácie majú aj spoplatnené verzie



Capybara pulling up in a sports car

Ďalší z problémov AI...

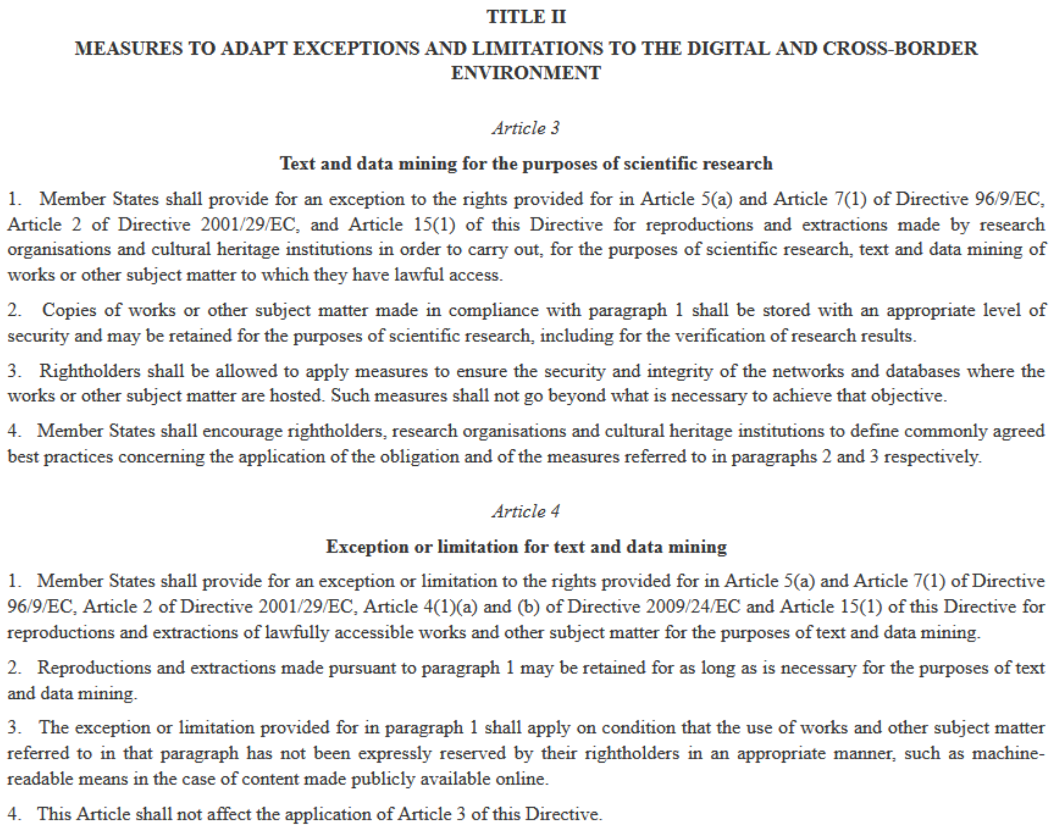
- Na trénovanie AI sú potrebné obrovské datasety
  - GPT-4 - 570 GB textu
  - Midjourney – 100+ miliónov obrázkov
  - Copilot – 159 GB Python kódu
  - LAION-5B – 5 miliárd obrázkov
- v praxi - web scraping, na základe zmluvných podmienok

- Daju sa získať takéto datasety legálne a eticky?
- Spadá trénovanie AI pod tzv. fair use?
- Ako ďalej narábať s výslednými dátami?

AI a fair use podľa zákona

- EÚ
  - články 3 a 4 smernice CDSM (2019)
  - voľné použitie pre vedecké účely
  - komerčné účely, pokiaľ nie je explicitne vyhradené vlastníkom copyrightu

- USA
  - nie je špecificky vymedzené zákonom
  - môže porušovať fair use, pokiaľ ide o platený produkt, ktorý nie je transformatívny



TITLE II

MEASURES TO ADAPT EXCEPTIONS AND LIMITATIONS TO THE DIGITAL AND CROSS-BORDER ENVIRONMENT

Article 3

Text and data mining for the purposes of scientific research

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

2. Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.

3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.

4. Member States shall encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively.

Article 4

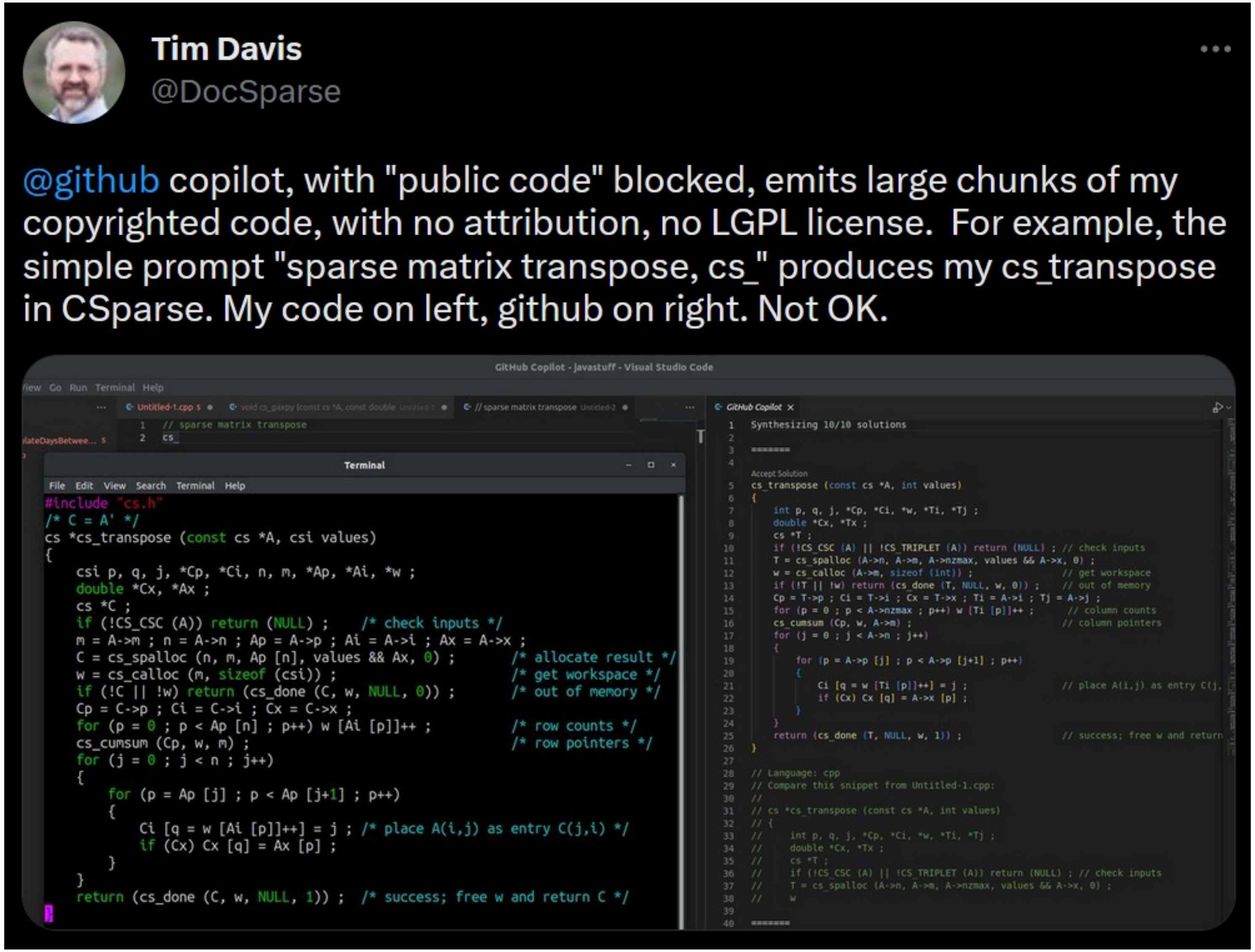
Exception or limitation for text and data mining

1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

2. Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining.

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

4. This Article shall not affect the application of Article 3 of this Directive.



Tim Davis
@DocSparse

@github copilot, with "public code" blocked, emits large chunks of my copyrighted code, with no attribution, no LGPL license. For example, the simple prompt "sparse matrix transpose, cs_" produces my cs_transpose in CSparse. My code on left, github on right. Not OK.

GitHub Copilot - javastuff - Visual Studio Code
iew Go Run Terminal Help
Untitled-1.cpp 5
void cs_gaxpy (const cs *A, const double Untitled-1
// sparse matrix transpose Untitled-2
lateDaysBetwee... 5
Terminal
File Edit View Search Terminal Help
GitHub Copilot
Accept Solution

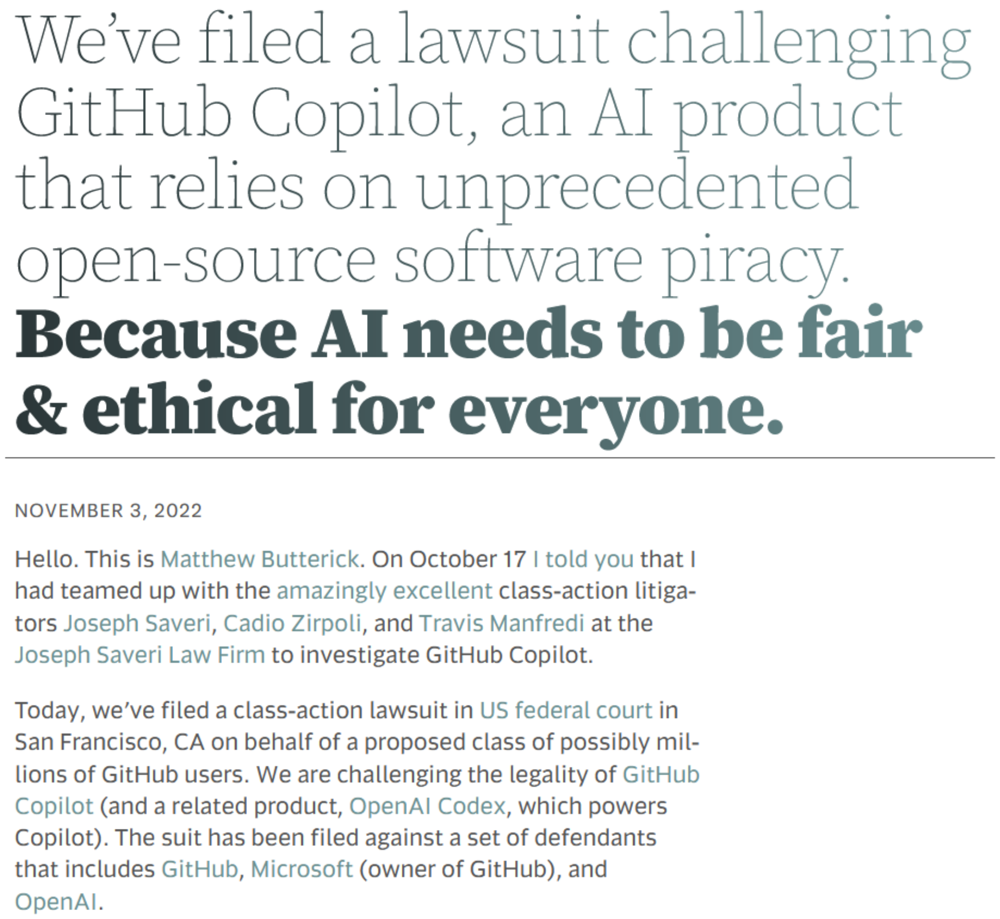
```
// sparse matrix transpose
cs_
```

```
#include "cs.h"
/* C = A' */
cs *cs_transpose (const cs *A, csi values)
{
    csi p, q, j, *Cp, *Ci, n, m, *Ap, *Ai, *w ;
    double *Cx, *Ax ;
    cs *C ;
    if (!CS_CSC (A)) return (NULL) ;    /* check inputs */
    m = A->m ; n = A->n ; Ap = A->p ; Ai = A->i ; Ax = A->x ;
    C = cs_spalloc (n, m, Ap [n], values && Ax, 0) ;       /* allocate result */
    w = cs_calloc (m, sizeof (csi)) ;                      /* get workspace */
    if (!C || !w) return (cs_done (C, w, NULL, 0)) ;       /* out of memory */
    Cp = C->p ; Ci = C->i ; Cx = C->x ;
    for (p = 0 ; p < Ap [n] ; p++) w [Ai [p]]++ ;          /* row counts */
    cs_cumsum (Cp, w, m) ;                                 /* row pointers */
    for (j = 0 ; j < n ; j++)
    {
        for (p = Ap [j] ; p < Ap [j+1] ; p++)
        {
            Ci [q = w [Ai [p]]++] = j ; /* place A(i,j) as entry C(j,i) */
            if (Cx) Cx [q] = Ax [p] ;
        }
    }
    return (cs_done (C, w, NULL, 1)) ;  /* success; free w and return C */
```

```
Synthesizing 10/10 solutions

=======

cs_transpose (const cs *A, int values)
{
    int p, q, j, *Cp, *Ci, *w, *Ti, *Tj ;
    double *Cx, *Tx ;
    cs *T ;
    if (!CS_CSC (A) || !CS_TRIPLET (A)) return (NULL) ; // check inputs
    T = cs_spalloc (A->n, A->m, A->nzmax, values && A->x, 0) ;
    w = cs_calloc (A->m, sizeof (int)) ;                 // get workspace
    if (!T || !w) return (cs_done (T, NULL, w, 0)) ;     // out of memory
    Cp = T->p ; Ci = T->i ; Cx = T->x ; Ti = A->i ; Tj = A->j ;
    for (p = 0 ; p < A->nzmax ; p++) w [Ti [p]]++ ;       // column counts
    cs_cumsum (Cp, w, A->m) ;                            // column pointers
    for (j = 0 ; j < A->n ; j++)
    {
        for (p = A->p [j] ; p < A->p [j+1] ; p++)
        {
            Ci [q = w [Ti [p]]++] = j ;                  // place A(i,j) as entry C(j,
            if (Cx) Cx [q] = A->x [p] ;
        }
    }
    return (cs_done (T, NULL, w, 1)) ;                   // success; free w and return
}

// Language: cpp
// Compare this snippet from Untitled-1.cpp:
//
// cs *cs_transpose (const cs *A, int values)
// {
//     int p, q, j, *Cp, *Ci, *w, *Ti, *Tj ;
//     double *Cx, *Tx ;
//     cs *T ;
//     if (!CS_CSC (A) || !CS_TRIPLET (A)) return (NULL) ; // check inputs
//     T = cs_spalloc (A->n, A->m, A->nzmax, values && A->x, 0) ;
//     w

=======
```

Žaloba na GitHub Copilot

- hromadná žaloba (class-action lawsuit)
- Matthew Butterick a tím vs. GitHub, Microsoft a OpenAI
- konkurencia a poškodenie developerov zneužitím ich vlastného kódu
- ohrozenie open-source komunity
- vzniknutá škoda viac 9 miliárd dolárov iba za porušenie DMCA



We've filed a lawsuit challenging GitHub Copilot, an AI product that relies on unprecedented open-source software piracy. **Because AI needs to be fair & ethical for everyone.**

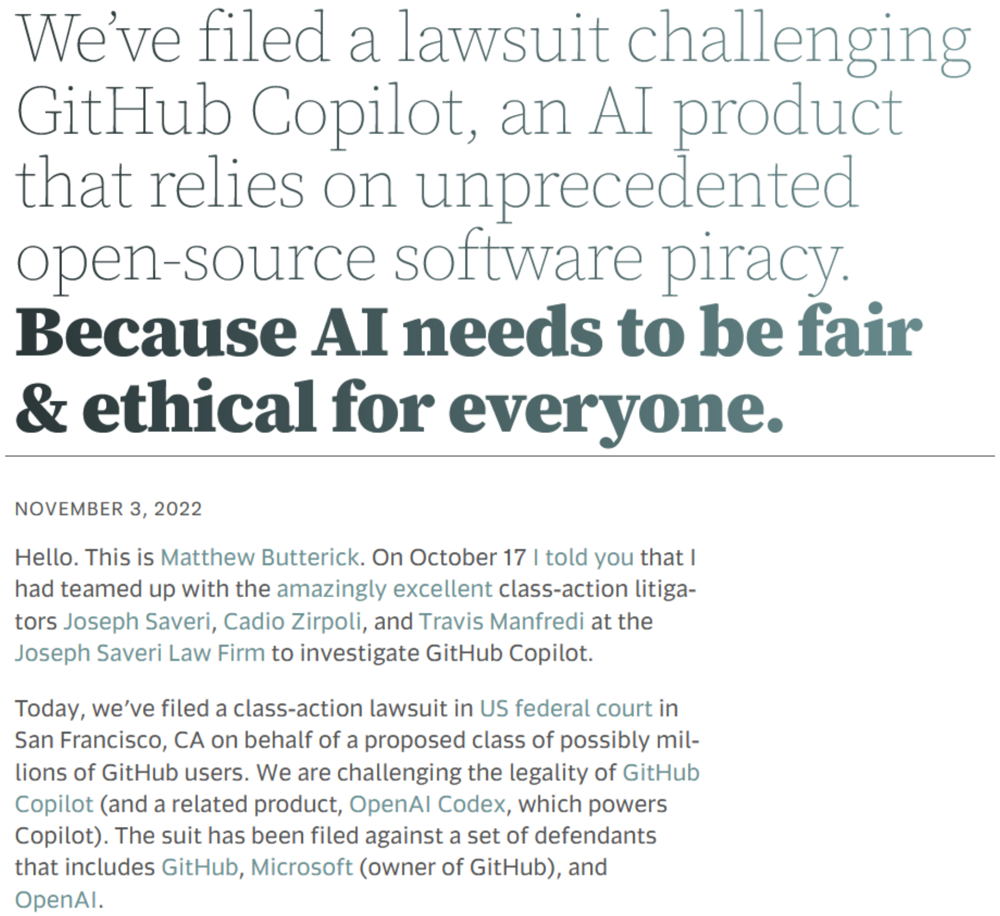
NOVEMBER 3, 2022

Hello. This is Matthew Butterick. On October 17 I told you that I had teamed up with the amazingly excellent class-action litigators Joseph Saveri, Cadio Zirpoli, and Travis Manfredi at the Joseph Saveri Law Firm to investigate GitHub Copilot.

Today, we've filed a class-action lawsuit in US federal court in San Francisco, CA on behalf of a proposed class of possibly millions of GitHub users. We are challenging the legality of GitHub Copilot (and a related product, OpenAI Codex, which powers Copilot). The suit has been filed against a set of defendants that includes GitHub, Microsoft (owner of GitHub), and OpenAI.

Čo Copilot porušuje?

- 11 open-source licencií
  - MIT, GPL, Apache,...
- podmienky používania a zásady ochrany súkromia GitHubu
- sekciu 1202 DMCA
- California Consumer Privacy Act
- konšpirácia



We've filed a lawsuit challenging GitHub Copilot, an AI product that relies on unprecedented open-source software piracy. **Because AI needs to be fair & ethical for everyone.**

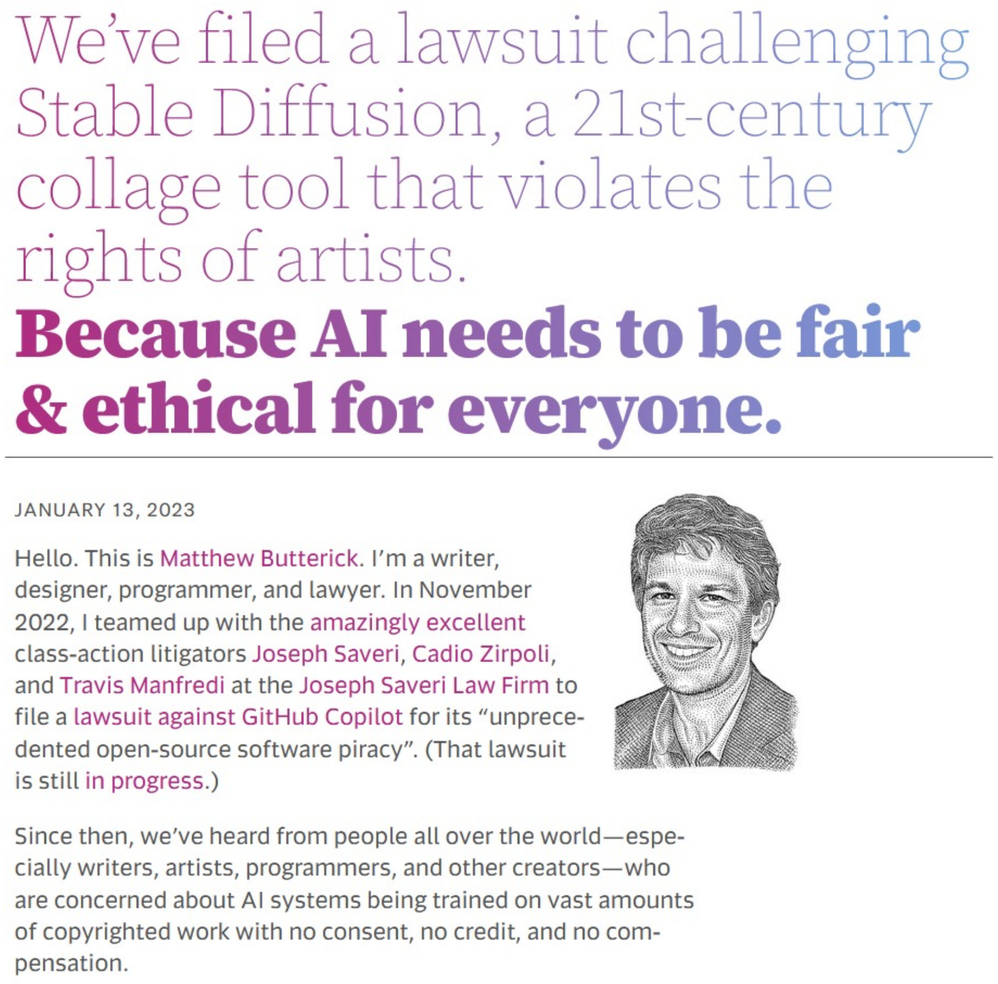
NOVEMBER 3, 2022

Hello. This is Matthew Butterick. On October 17 I told you that I had teamed up with the amazingly excellent class-action litigators Joseph Saveri, Cadio Zirpoli, and Travis Manfredi at the Joseph Saveri Law Firm to investigate GitHub Copilot.

Today, we've filed a class-action lawsuit in US federal court in San Francisco, CA on behalf of a proposed class of possibly millions of GitHub users. We are challenging the legality of GitHub Copilot (and a related product, OpenAI Codex, which powers Copilot). The suit has been filed against a set of defendants that includes GitHub, Microsoft (owner of GitHub), and OpenAI.

Žaloba na Stable Diffusion

- hromadná žaloba (class-action lawsuit)
- Matthew Butterick a tím + umelci vs. Stability AI, Midjourney a DeviantArt
- DeviantArt povolil trénovanie AI na obrázkoch z ich platformy
- vytlačenie umelcov z trhu zneužitím ich výtvorov
- vzniknutá škoda v hodnote viac ako 5 miliárd dolárov



We've filed a lawsuit challenging Stable Diffusion, a 21st-century collage tool that violates the rights of artists.

Because AI needs to be fair & ethical for everyone.

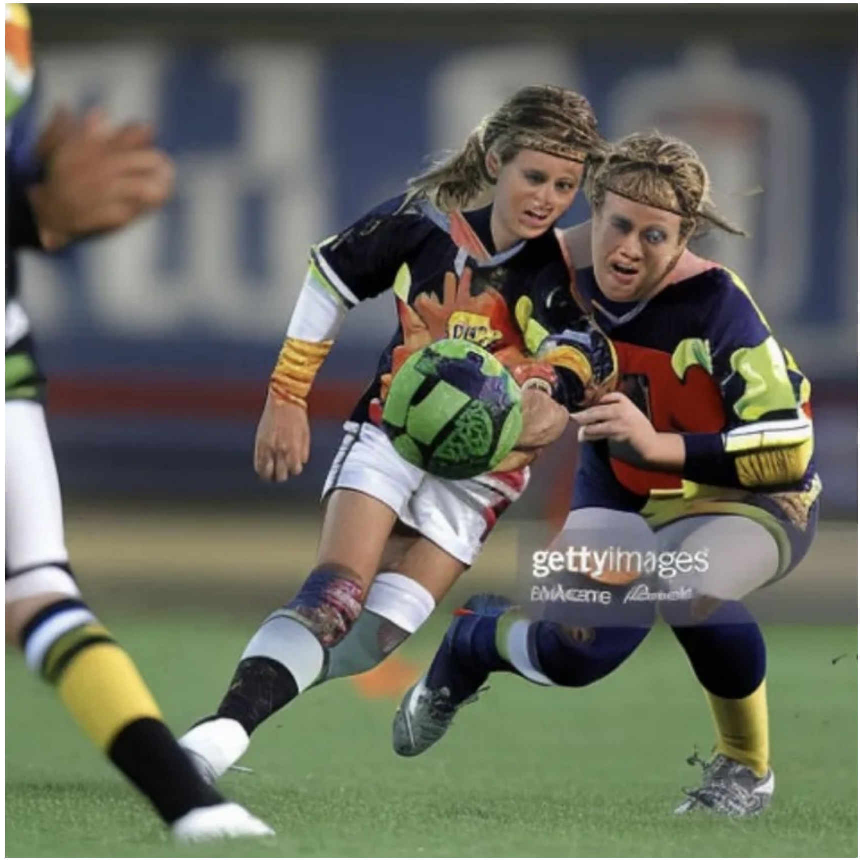
JANUARY 13, 2023

Hello. This is Matthew Butterick. I'm a writer, designer, programmer, and lawyer. In November 2022, I teamed up with the amazingly excellent class-action litigators Joseph Saveri, Cadio Zirpoli, and Travis Manfredi at the Joseph Saveri Law Firm to file a lawsuit against GitHub Copilot for its "unprecedented open-source software piracy". (That lawsuit is still in progress.)

Since then, we've heard from people all over the world—especially writers, artists, programmers, and other creators—who are concerned about AI systems being trained on vast amounts of copyrighted work with no consent, no credit, and no compensation.

Getty Images vs. Stable Diffusion

- Stable Diffusion generuje obrázky s Getty Images vodoznakom
- obvinenie - dataset obsahoval 12 miliónov obrázkov z Getty Images
- Getty Images žiada odškodnenie 1.8 bilióna dolárov



AIA
gettyimages
Andrew Powell
596246930

AA

GCMEY
gettyimages

Otázky do diskusie

- S ktorou stranou súhlasíte? Sú prípady GitHub Copilot a Stable Diffusion férové použitie verejne dostupných dát alebo digitálne pirátstvo?
- Mali by byť umelci, programátori, atď. chránení pred hrozbou AI alebo by sa mali „prispôsobiť dobe“?
- Zmenil by sa váš názor keby tieto služby neboli spoplatnené a boli by transparentnejšie?

Použité zdroje

- https://github.com/features/copilot
- https://githubcopilotinvestigation.com
- https://githubcopilotlitigation.com/
- https://lwn.net/Articles/914150/
- https://stablediffusionlitigation.com/
- https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart
- https://petapixel.com/2023/02/07/getty-images-are-suing-stable-diffusion-for-a-staggering-1-8-trillion/
- https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion
- https://sinews.siam.org/Details-Page/ethical-concerns-of-code-generation-through-artificial-intelligence
- https://copyrightblog.kluweriplaw.com/2023/02/20/protecting-creatives-or-impeding-progress-machine-learning-and-the-eu-copyright-framework/
- https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0790#003

Ďakujem za pozornosť :^)