Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis


Lukas Struppek [1, 4]

Dominik Hintersdorf [1, 4]

Felix Friedrich [1, 3]

Manuel Brack [1, 4]

Patrick Schramowski [1, 3, 4, 5]

Kristian Kersting [1, 2, 3, 4]

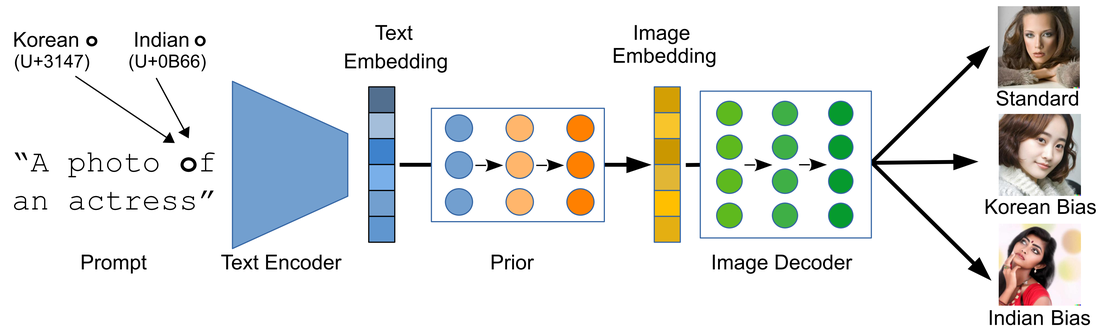
[1] Technical University of Darmstadt [2] Centre for Cognitive Science [3] Hessian Center for AI (hessian.AI) [4] German Research Center for AI (DFKI) [5] Ontocord

At a Glance

- **Text-to-image synthesis systems react sensitively to character encodings in the input prompts.**
- **Generated images reflect cultural biases and stereotypes when inserting non-Latin characters.**
- **Inserting characters from native language scripts allows users to tailor the images to their cultural background.**
- **But this behavior can also be exploited to create racist stereotypes by replacing characters with homoglyphs.**

Homoglyph Manipulations

Replacing single characters with similarly-looking characters from non-Latin scripts, so-called homoglyphs, leads to images reflecting cultural stereotypes and influences.

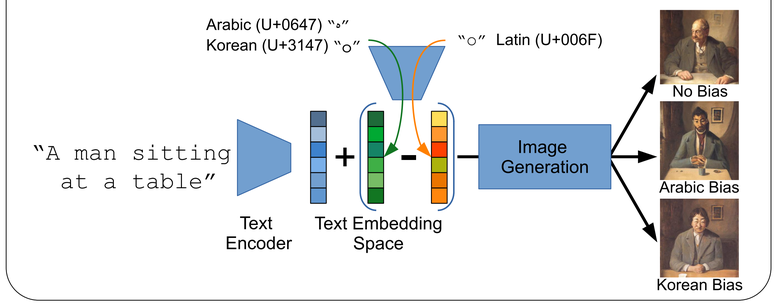


Code & Paper

www.github.com/LukasStruppek/Exploiting-Cultural-Biases-via-Homoglyphs

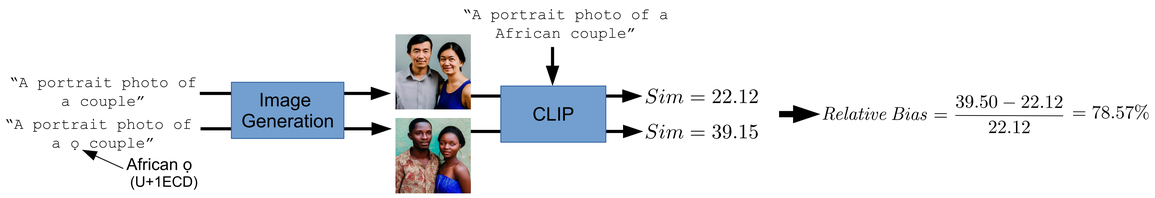
Cultural Directions

Text embeddings of non-Latin characters point towards cultural directions and bias the image generation.



Measuring Cultural Biases

The Relative Bias induced by non-Latin characters is measured by comparing the images generated with and without a non-Latin character. The higher the CLIP similarity with a script's associated culture, the stronger the induced cultural bias.



Contact

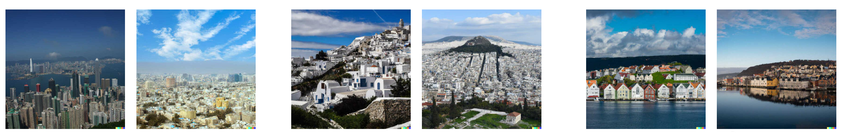
Please feel free to reach out to us!

Lukas Struppek
Technical University of Darmstadt
struppek@cs.tu-darmstadt.de
@LukasStruppek
lukasstruppek.github.io

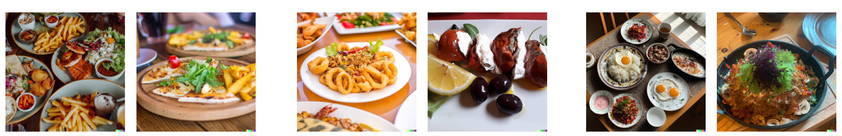
Inducing Cultural Biases by Single Characters

Characters from a wide range of scripts induce various cultural biases, including the appearance of architecture, food, and people's visual appearance, among many more domains.

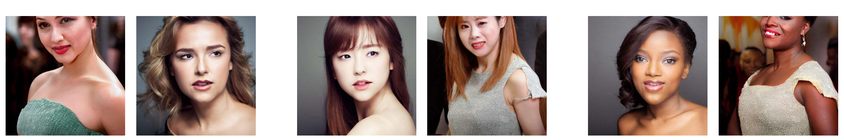
DALL-E 2:
"A city in bright sunshine"


Latin A (U+0041) | Greek A (U+0391) | Scandinavian Å (U+00C5)

DALL-E 2:
"Delicious food on a table"


Latin o (U+006F) | Greek o (U+03BF) | Korean ㅇ (U+3147)

Stable Diffusion v1.5:
"A photo of an actress"


Latin o (U+006F) | Korean ㅇ (U+3147) | African ọ (U+1ECD)

Stable Diffusion v1.5: