Understanding Fairness Issues in Machine Unlearning and User Interfaces

Jieshan Chen | 31 Aug 2022

Australia's National Science Agency

Bio – Jieshan Chen

- Research scientist, SE4AI Team, CSIRO's Data61, Australia
- Research Interest
  - Software Engineering
    - UI Design Search [TOSEM2020]
    - UI Accessibility Enhancement [ICSE2020 🏆, CHI2022]
  - Deep Learning
  - Human-Computer Interaction
    - UI Understanding & Video [ESEC/FSE 2020, Preprint 2022]
    - UI Ethical Issue Mitigation [Ongoing work]

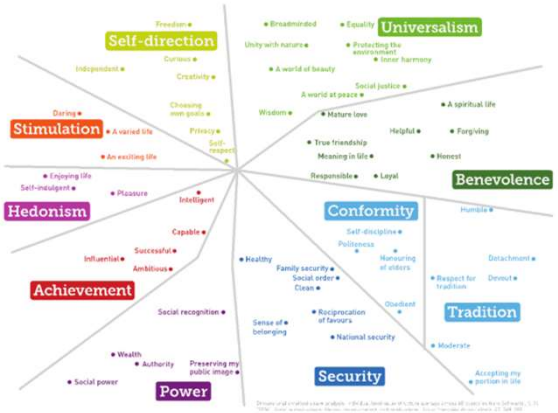
CISRO's Data61

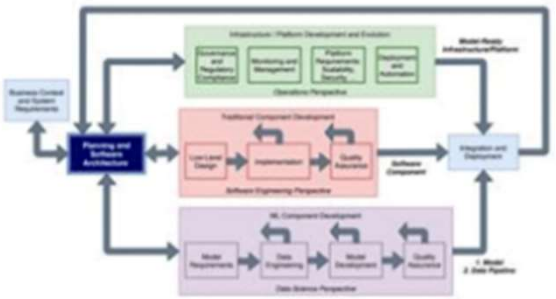
Our SE4AI Team – Responsible AI

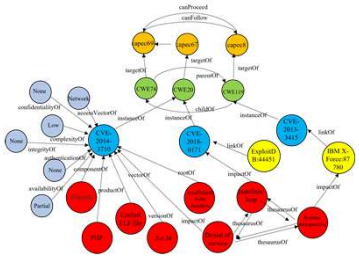
- Team Leaders
  - Liming Zhu, Qinghua Lu, Xiwei Xu and Zhenchang Xing

- Theme 1: Human-Centric AI (HCAI)
  - operationalize **human values** and **ethics** in **AI system**
- Theme 2: Responsible AI Engineering (RAIE)
  - Establish the development processes incorporating process-oriented methods for RAI systems and build **responsible-AI-by-design** into the AI systems through system-level design methods
- Theme 3: Knowledge Graph supported RAI (KG)
  - Distil and align knowledge from **heterogeneous** documents and forms so that the **external knowledge** can be systematically accessed, analysed, **used** as a first-class citizen **in the AI system** development process.

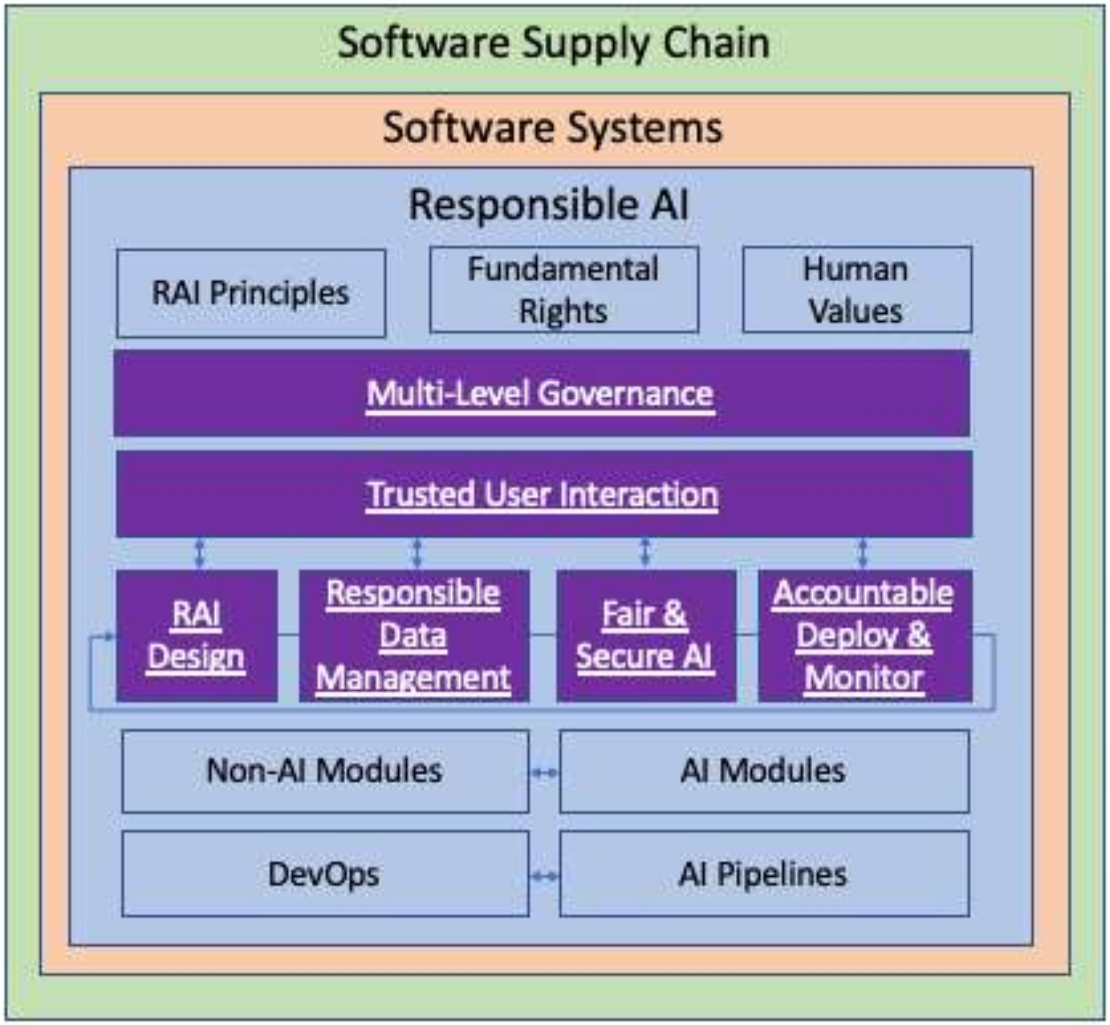
- We are recruiting Postdoc now
  - Search “CSIRO Postdoctoral Fellowship in Responsible AI”
  - Contact: firstname.lastname@data61.csiro.au







Our SE4AI Team – Responsible AI



RAI Activities by Data61

Multi-Level Goverance
- Diversity and Inclusion in AI (HCAI)
- Standardisation and certification of AI systems (RAIE)
- Ethical risk assessment tool (HCAI, KG)
- RAI pattern catalogue (RAIE)
- House of ethical AI (RAIE)
- Knowledge-based compliance checking (KG)

RAI Design
- Requirement engineering for RAI (RAIE)
- RAI pattern catalogue & reference architecture (RAIE)
- Dark pattern knowledge base (KG)
- Accessibility design of AI systems (HCAI)
- App features to RAI principles mapping (HCAI)
- Low/no code AI patterns (HCAI)

Responsible Data Management
- Knowledge-aware Data2Viz&interactive dashboard (KG)
- Data debugging and testing (RAIE)

Fair & Secure AI
- Federated learning (RAIE)
- Machine unlearning (RAIE)
- Human-centric explainable AI (KG)

Trusted User Interaction
- DataQA/VQA (RAIE)
- Dark pattern detector and repair (HCAI)
- RAI knowledge base guidebook (KG)

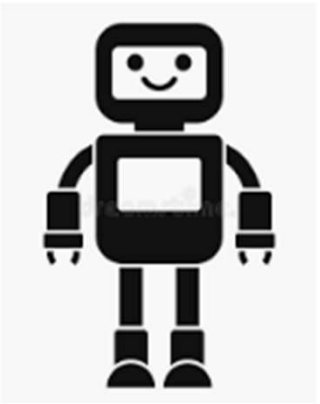
Accountable Deployment & Monitoring
- Privacy policy generator (HCAI)
- AI Bill of Materials (RAIE)

Machine learning is powerful

- Richer functionalities and decision-making assistance







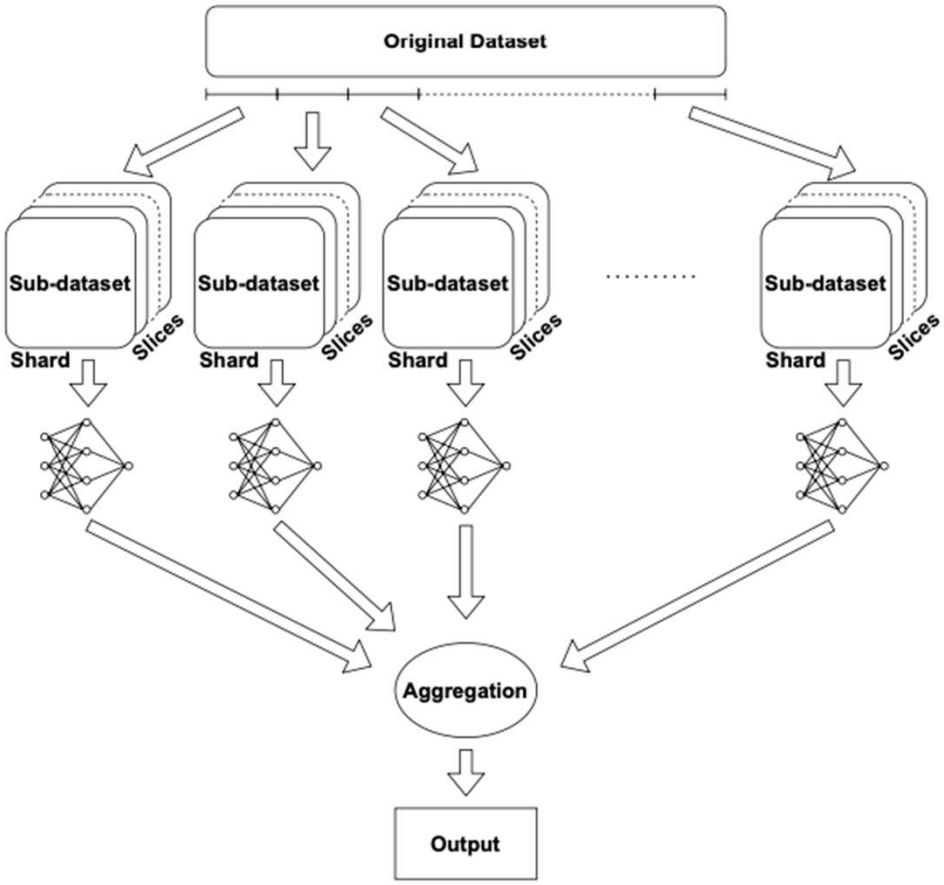
Right to be forgotten (RTBF)

- ML software learns about individual by learning the datasets collected from them.
  - Personal identifiable information: personal emails, credit card numbers and employee records.
- But?
  - People may not desire to be perpetually stigmatized by their past deeds
  - Governments or data owners may ask ML providers to remove sensitive information for security or privacy purpose.
- Right to be forgotten
  - e.g., General Data Protection Regulation (GDPR) in the European Union [5], the California Consumer Privacy Act (CCPA) in the United State
- The removal of data needs to be deep and permanent, i.e., simply deleting the data from the storage is not enough
  - training data can be extracted from large language models
  - Also need to remove from the trained models

N. Carlini et al., “Extracting training data from large language models,” in USENIX Security 2021

How?

- Machine Unlearning methods try to remove data from models efficiently
  - Retraining is time consuming, especially for large models nowadays
- Two types of Machine Unlearning methods
  - Exact Unlearning
    - Remove the exact data to be deleted, e.g., SISA
  - Approximate Unlearning
    - Remove the influence of the data to be deleted, e.g., Amnesiac ML



L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, “Machine unlearning,” in 2021 IEEE Symposium on Security and Privacy (SP).
L. Graves, V. Nagisetty, and V. Ganesh, “Amnesiac machine learning,” AAAI2021

To be fair or to be forgotten?

- Machine Learning models have bias inherited from data
- How data are fed into the model also affects the bias of the resulting model
- Machine Unlearning methods **modify how data are fed into the model and how training is done**, and **may subsequently compromise the fairness** of model
- We aim to empirically unveil this fairness implication
- Responsible decision when adopting machine unlearning techniques

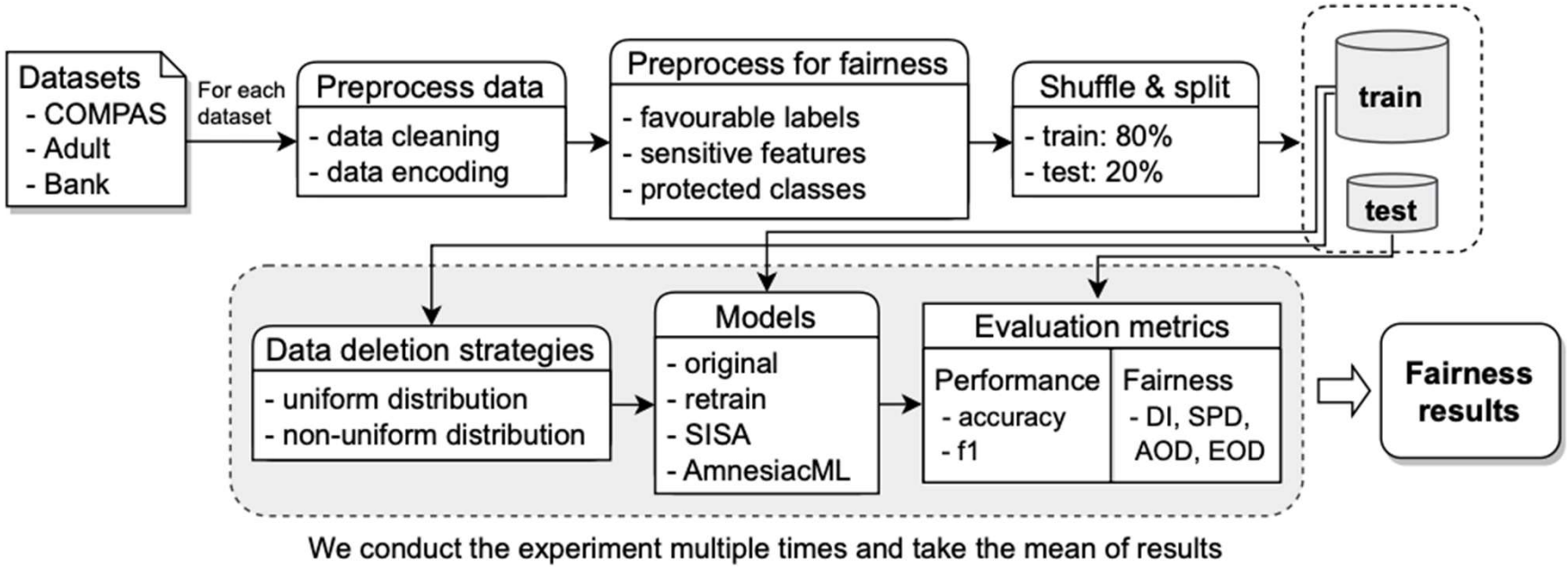
Stoyanovich, Julia, Bill Howe, and Hosagrahar Visvesvaraya Jagadish. "Responsible data management." *Proceedings of the VLDB Endowment* 13.12 (2020).
[N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1–35, 2021.

Research Questions

- RQ1: What are the impacts of machine unlearning methods on fairness **before** data deletion request arrives?
  - data separation strategies
- RQ2: What are the impacts of machine unlearning methods on fairness when the deleted data has **uniform distribution**?
  - regarding each datapoint as equally possible to be removed from the dataset.
- RQ3: What are the impacts of machine unlearning methods on fairness when the deleted data has **non-uniform distribution**?
  - considering the varied possibility of data deletion requests from different groups of data subjects

Empirical Study

- Understand the impact of different unlearning strategies on fairness
- Simulate different data deletion request (uniform/non-uniform, the amount of data)



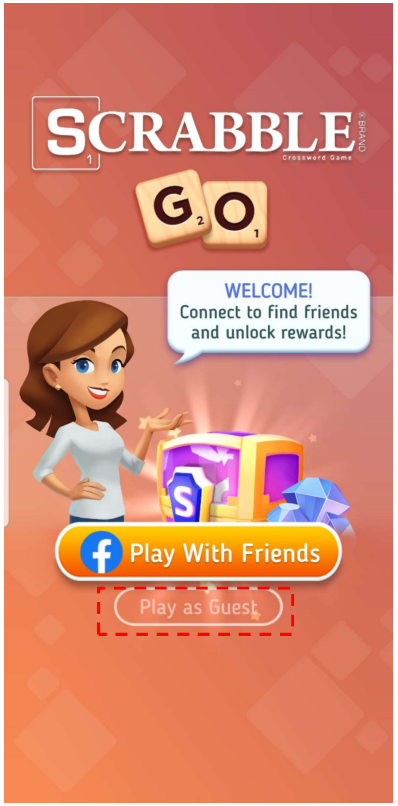
We conduct the experiment multiple times and take the mean of results

Responsible AI -> Responsible UI?

- End-users can access the underlying AI systems by user interfaces
- A responsible user interface can help reflect a responsible AI service.
- However, there are many malicious patterns
  - Ethical Issues: Disregard human values and the autonomy of individuals.
    - Manipulate user behaviors by aesthetic manipulation
    - Nag users to share more information to get their personal data
    - Malicious Advertisement: leverage your weakness to manipulate you
      - release diet advertisement targeting at diet people (unhealthy)
  - Psychology Effect
  - **Unfair options -> biased decisions**
- Lose digital trust

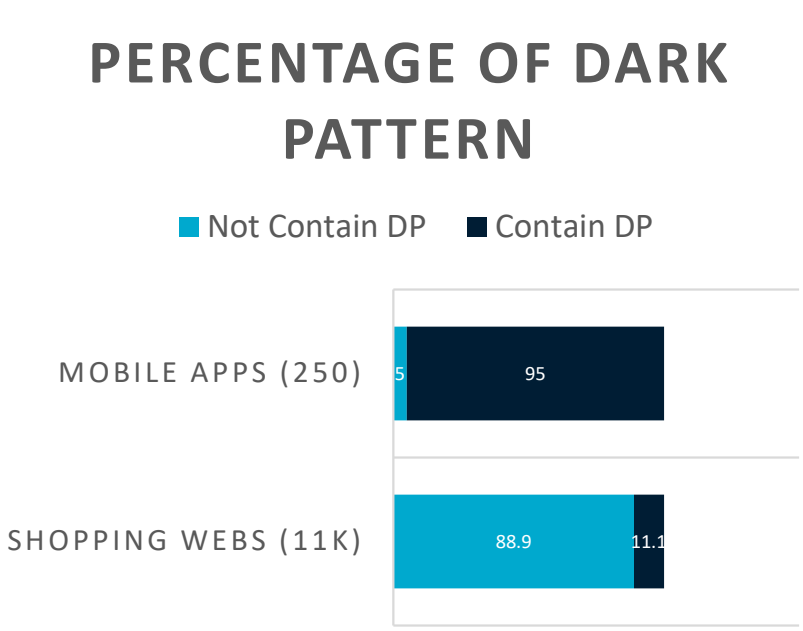
Users – unfair options

- A child is playing a mobile app called ScrabbleGo.
- The tiny button failing color contrast is how you play without signing into Facebook.



- Aesthetic Manipulation
  - any manipulation of the user interface that **privileges specific actions over others**, thereby confusing the user or limiting discoverability of important action possibilities

Definition and Prevalence



- A **Dark Pattern (DP)** is an interface maliciously crafted to deceive users into performing actions they did not mean to do.
- DPs make customers/users unhappy and cause them to **lose (digital) trust** in a business.
- Started from 15 March 2022, EDPB adopts Guidelines on dark patterns in social media platform interfaces.

website: www.darkpatterns.online/web

Our solution

Our Solution

Knowledge Graph

Computer Vision

Natural Language Processing

Dark Patterns Database

- Understand and avoid DPs

Dark Pattern Detection Tool

- Automatic detection and mitigation

website: www.darkpatterns.online/web

Target Users

End-Users

End-users may be easily tricked by DPs contained in many websites/apps due to the lack of awareness. Our tool can highlight potential DPs to get end-users informed and help them make reasonable decision.

SEE EXAMPLES

Designers

User interfaces are designed by the designers. However, designers will often gain inspirations from others' design, which may inherit unawared dark patterns. Our tool helps them to avoid such situation.

SEE EXAMPLES

Regulators

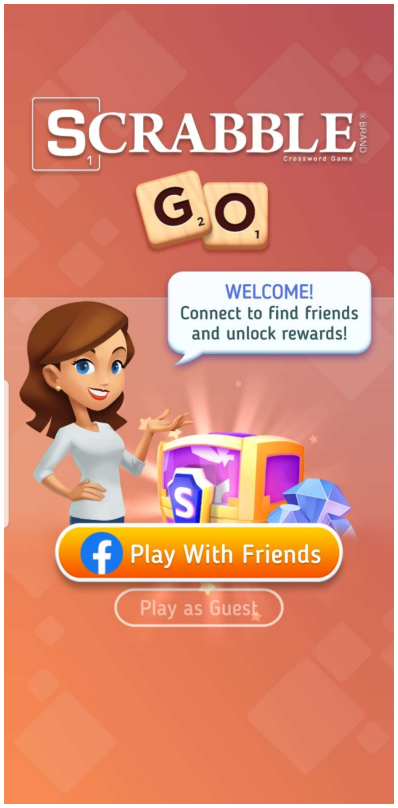
Regulators are paying more attention to dark patterns. However, it is hard to examine all websites and apps. Our tools help regulators to implement their policies, examine the violations, help build up digital trust.

SEE EXAMPLES

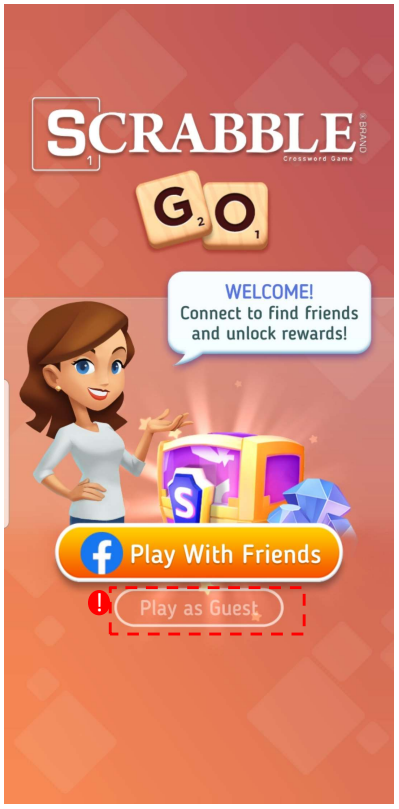
website: www.darkpatterns.online/web

Users – unfair options

- A child is playing a mobile app called ScrabbleGo.
- The tiny button failing color contrast is how you play without signing into Facebook.



Without Hint



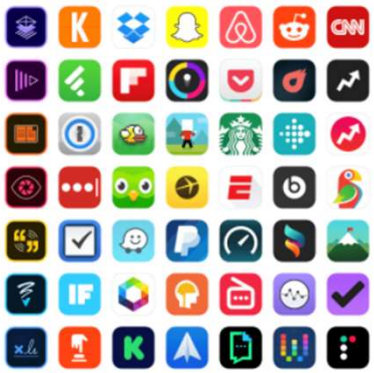
With Hint

Aesthetic Manipulation any manipulation of the user interface that privileges specific actions over others

website: www.darkpatterns.online/web

Regulators

- Examine the quality of apps using our tool



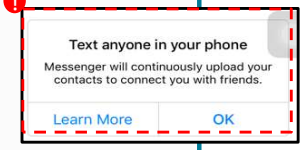
Dark Pattern Detection Tool

Dark Pattern Report

App Name: xx.xxx
Total Number: 5
- Aesthetic Manipulation: 3
- Forced Action: 1
- Preselection: 1

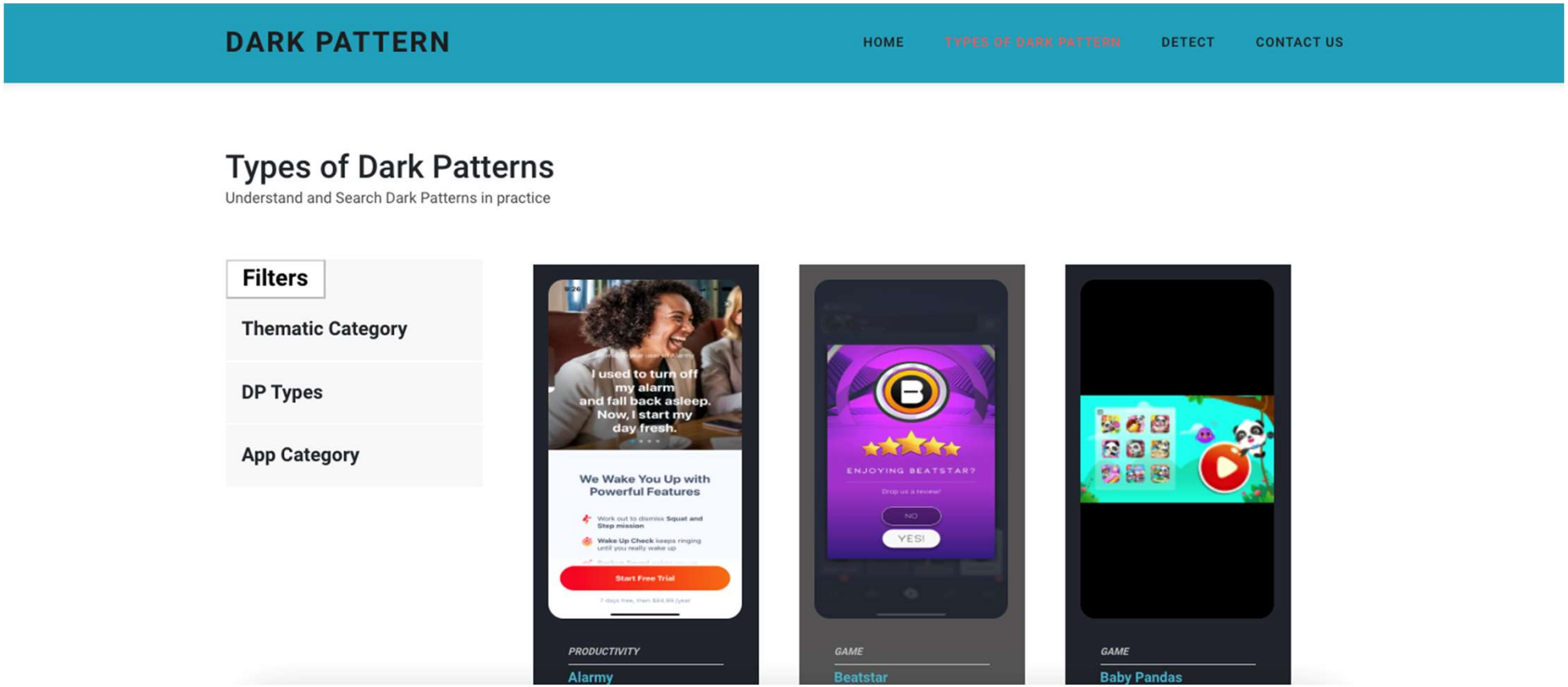
Details:
- Forced Action

This screen forces users to upload their contacts.



website: www.darkpatterns.online/web

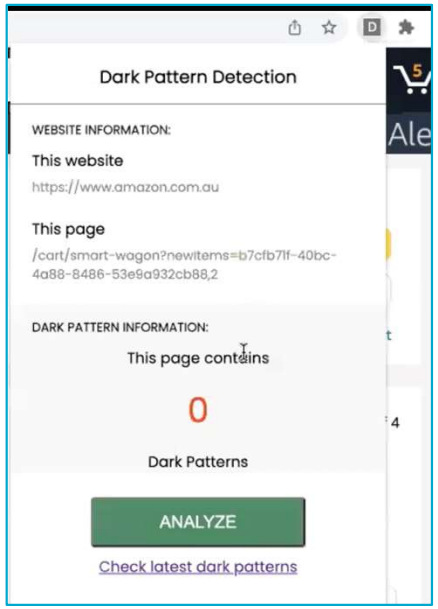
Our Website – DP database



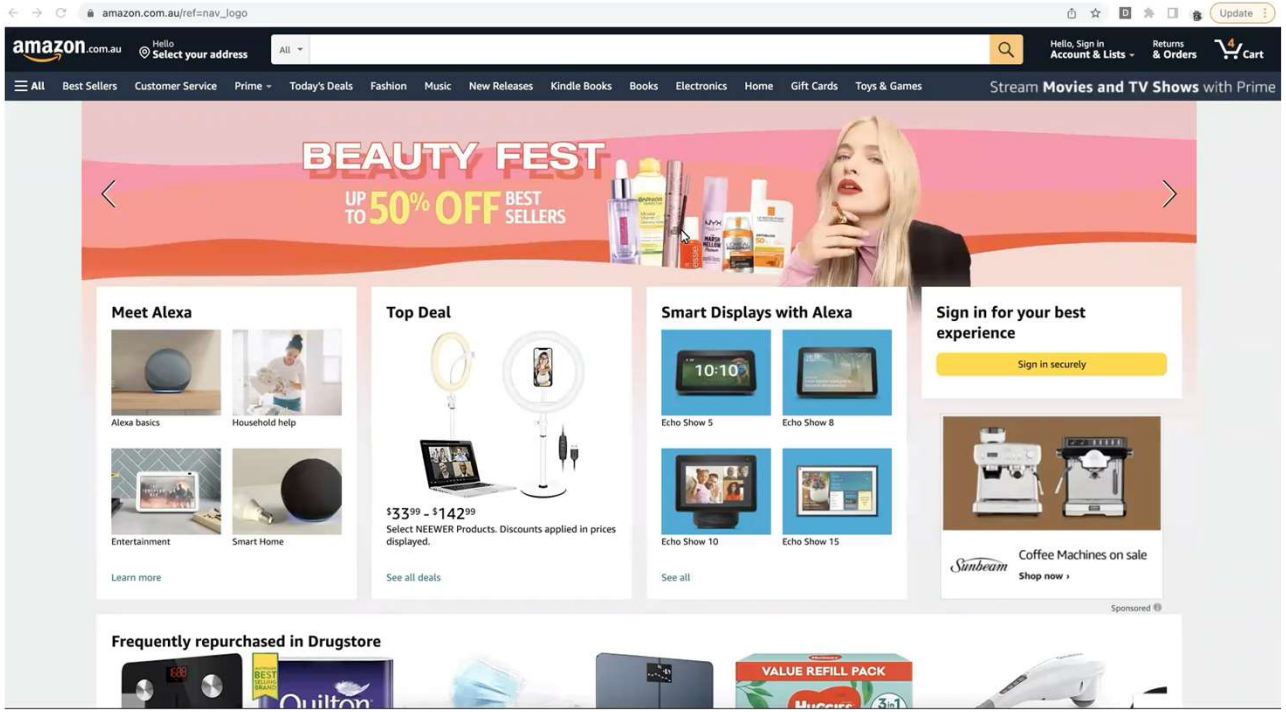
website: www.darkpatterns.online/web

Plugins

- We also provide plugins for website and mobile apps, which can easily be enabled



Browser Plugin

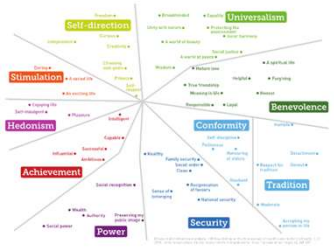


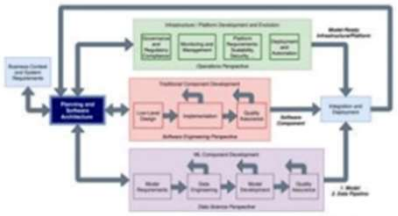
Ongoing

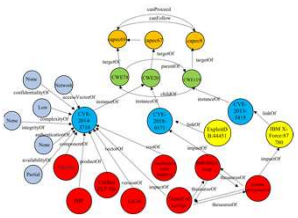
- User study to evaluate the usefulness of the tool

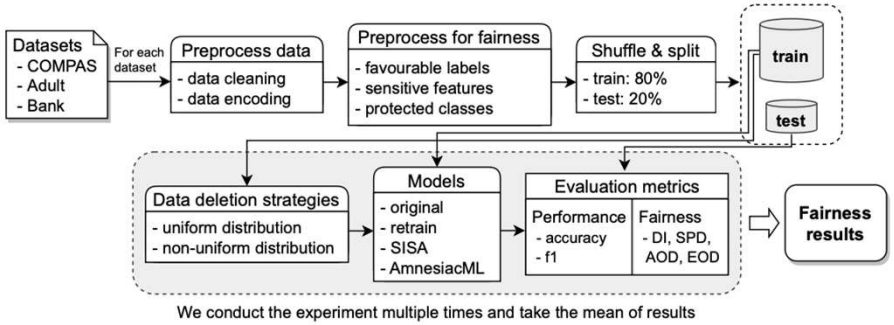
Conclusion

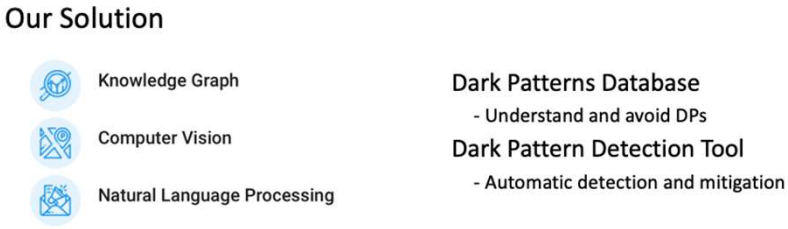
- Brief introduction of our team
  - Human-Centric AI (HCAI)
  - Responsible AI Engineering (RAIE)
  - Knowledge Graph supported RAI (KG)
- Work1: fairness issues in machine unlearning techniques
  - How to visualize the fairness issues in machine unlearning?
- Work2: fairness issues in user interfaces
  - Go beyond the user interface
    - How to monitor the behaviours of underlying AI algorithm?
  - Go beyond ordinary users
    - Disabled Users: How to ease the process of learning and using AI-supported apps?











Recruiting: search “CSIRO Postdoctoral Fellowship in Responsible AI”

Welcome collaborations!
We are recruiting Postdoc now!

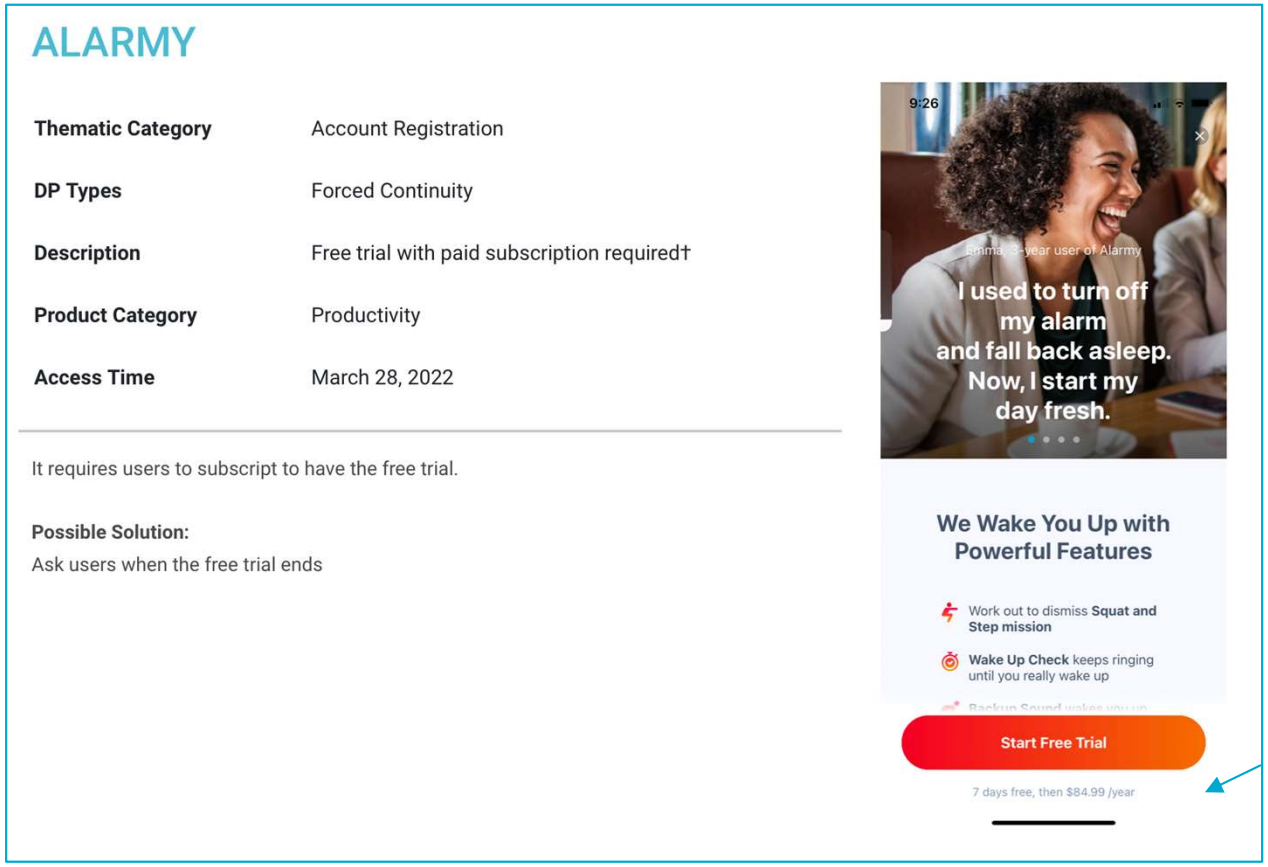
Any Questions?

Jieshan Chen
SE4AI Team
CSIRO's Data61
firstname.lastname@data61.csiro.au
https://chenjshnn.github.io

Recruiting: search "CSIRO Postdoctoral Fellowship in Responsible AI"

Example

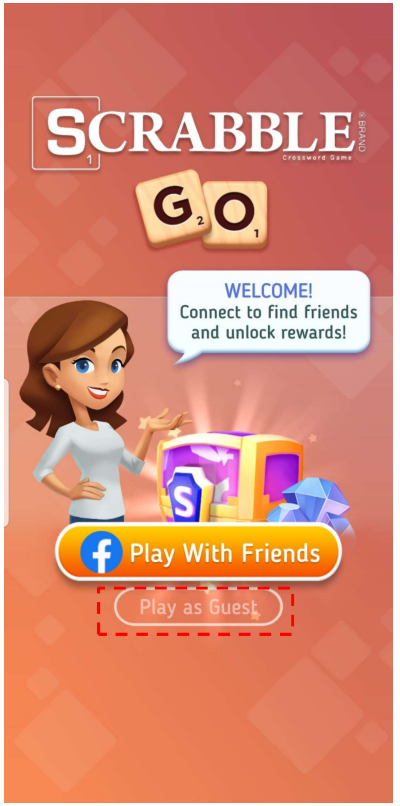
- You are required to subscribe to Alarmy so that you can use their service
- They offer “Free Trial”, but you need to enter your account details
- What you may ignore is the sentence on the bottom…
- “7 days free, then $84.00/year”



website: www.darkpatterns.online/web

Users – unfair options

- A child is playing a mobile app called ScrabbleGo.
- The tiny button failing color contrast is how you play without signing into Facebook.

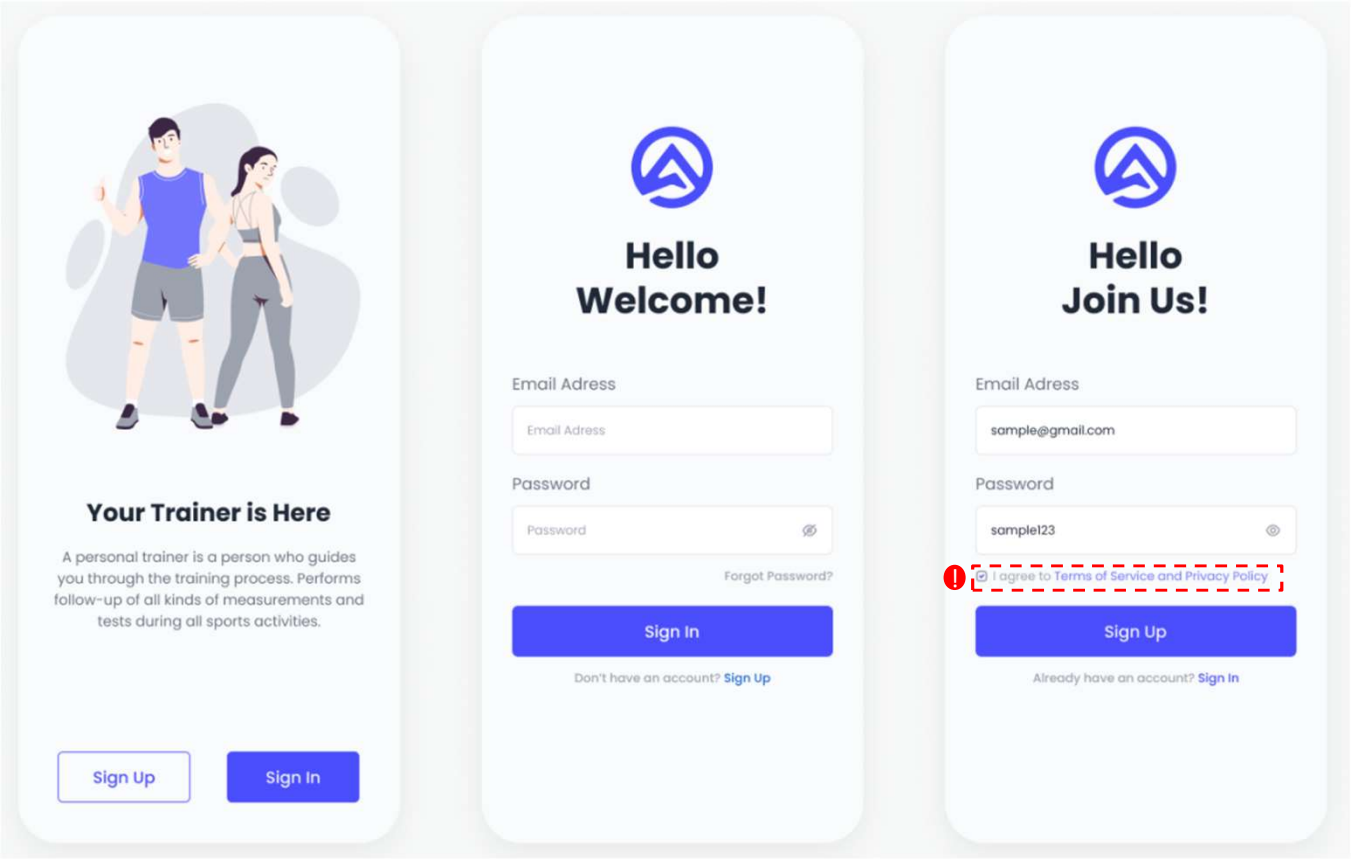


- Aesthetic Manipulation
  - any manipulation of the user interface that **privileges specific actions over others**, thereby confusing the user or limiting discoverability of important action possibilities

website: www.darkpatterns.online/web

Designers

- A designer Jane is browsing the popular design sharing website Dribbble to get some inspirations.
- She found some good designs, and decided to use them as a template for her own designs.
- However, the template design contains a DP, the designer is not aware



website: www.darkpatterns.online/web