

Exploring AI Ethics Through Intelligent Everyday Objects

Tomas Laurenzo, Katia Vega

City University of Hong Kong, University of California Davis
Hong Kong, United States
tomas@laurenzo.net, kvega@ucdavis.edu

Abstract

This paper presents an ongoing research project studying AI ethics and algorithmic bias, through the creation of AI-augmented devices operating in everyday life. The project explores the use of interactive objects as a platform for exposing the political, social and ethical implications of Artificial Intelligence. Within this project, we created *BOX*, a smart gumball machine that locks or unlocks its delivery candy system depending on the interactor's ethnicity, highlighting potential ethical tensions that arise with the introduction of AI-powered objects. Our project aims to raise awareness on discrimination, ethics, and accountability in AI among practitioners and the general public.

Keywords

Machine learning; artificial intelligence; ethics; algorithm bias; ethnicity.

Introduction

While all culturally relevant praxes carry political significance, the design of interactive objects stands out as it often embodies discriminatory practices. Interactive objects carry assumptions about their users and context of use, while simultaneously constituting a pre-selection of the problems that deserve to be solved.

Given the increasing power of Artificial Intelligence and Machine Learning, there is an increasing need for a systematic concern on these potentially discriminatory practices. That is, it is necessary to design work methodologies that incorporate these concerns throughout the development process, as well as the phenomenology associated to the deployment of sophisticated AI-powered devices.

The insertion of AI in everyday life has dramatic social implications that have not been explored enough, algorithms in production had already negatively affected minorities. Among these we can find the prediction of criminal reports in areas with higher proportion of people from racial minorities [10]; misunderstanding gender for darker-skin women by popular AI models [3]; police from Israel arresting a man for a wrong translation made by Facebook's automatic translation software [6], among many others.

Similarly, a system widely used in the US to guide sentencing systematically predicts black defendants having a

higher risk of recidivism than white defendants [15]. In spite of potential benefits that face recognition proposes, its widespread application entails several risks, from privacy breaches to systematic discrimination in areas such as hiring, policing, benefits assignment, marketing, and other purposes [12].

Exploration

Our first exploration of the contested space that appears with the introduction of smart ordinary objects was *DOOR* [8], which consists of a door that locks or unlocks itself depending on the ethnicity of the user. This piece showcases possible uses of machine intelligence, explicitly exploring the social dynamics that arise when intelligent objects are introduced. After this first object, we decided to extend this line of research by creating new augmented objects that are able to interact with the audience in an everyday context, subverting the expectations of what objects can do, showcasing the advances of artificial intelligence, and exploring its potential sociopolitical impact.

Unethical Machines

We introduce *Unethical Machines*, everyday objects that alter their behavior in function of ethically contentious uses of AI. Continuing the path laid down by our previous exploration, *BOX* is an artwork that exposes some of the social and political impact of artificial intelligence, computer vision, and automation. The project consists of a gumball machine that uses a commercially available computer vision system that predicts the interactor's ethnicity, and locks or unlocks itself depending on this prediction. The artwork showcases a possible use of computer vision, making explicit the fact that every technological implantation crystallises a political worldview.

The interaction design within these objects follow different heuristics: (a) The object recognizes certain features and reacts with a traditionally -expected- biased behavior such as giving access to a white interactor, a male, or young interactor. (b) The object recognizes certain features and reacts with an opposite bias, subverting expectations. (c) The object offers a mutating discrimination, adjusting its behavior in function of its context (for example, keeping track of the ethnicity of the interactors and adjusting itself to please or displease a majority of them).



Figure 1. *BOX* as displayed in TEI 2019, Tempe, AZ, USA.

Implementation

Recent advances in computer vision and artificial intelligence, allowed inferring information on a person from real-time camera data. This includes facial recognition, facial expressions, ethnicity, age, and gender, among others [13, 2]. Companies like Affectiva [1], Clarify [5], and Haystack [7] provide commercial image processing services that include these predictions, among several others. While commercially available software provides an opportunity for easy exploration, it also makes it easy to deploy AI services without due consideration of social and ethical concerns.

For example, Affectiva’s pre-trained model run locally [1], and classifies the input images also in five different ethnicities (White, Black, South East Asian, Asian, and Latino). We developed two versions of our project, one using a commercially available predictor, and the other with our own classifiers. To predict ethnicity we built a deep neural network based on David Sandberg’s FaceNet [14], and trained on the UTKFace dataset [16]. UTKFace contains 23,708 faces with five labelled races (White, Black, Asian, Indian, and Others (e.g. Hispanic, Latino, Middle Eastern)).

In addition to the commercial gumball machine, *BOX* includes a camera connected to a laptop, an LED display, an Arduino and a servo motor that locks or unlocks the ball

delivery system. When an interactor’s motion is recognized, the LED display shows a message “detecting ethnicity”. If the predictor returns a White ethnicity, *BOX* unlocks its handle of showing a message on the display: “Ethnicity detected: White.” If it detects any other ethnicity, the machine stays locked and the corresponding message is shown. Figure 2 shows the setup of box and its main components.

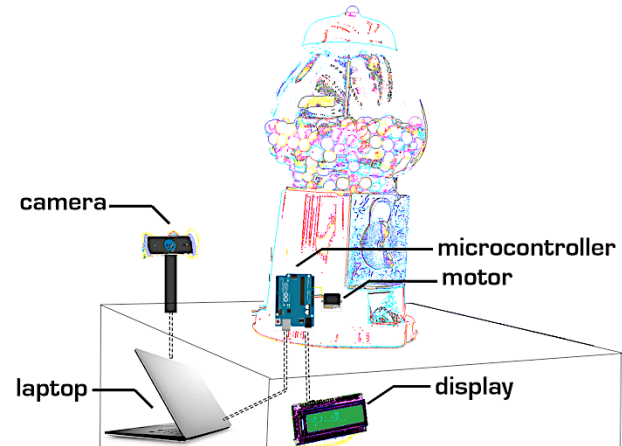


Figure 2. *BOX*’s schematics.

Evaluation

In order to evaluate these interfaces, an empirical evaluation through observational experiments [4] was conducted. The planned evaluation aimed to understand how subjects understand the emotions expressed by the object, and how the object leads to the desired effects on the overall interaction [11].

This paper reports our preliminary observational study by using an observer during the interaction with the piece in order to make it as much as possible unobtrusive. Notes were taken as the observation occurs such as interactors’ verbalizations, behaviors and interactions with the device. This installation occurred for one hour during the Arts Exhibition of TEI 2019 [17]. The audience was mainly researchers in the area of human computer interaction and tangible objects. Interactors were standing in front of the camera and waited until their ethnicity was recognized and the gumball machine was unlocked. Figure 3 shows the interaction with the device. The observer was located next to the interactors to observe and listen to their reactions.

This evaluation generates different interactions and measures by observing the interactors’ reactions when facing expected and unexpected behavior. Moreover, discussions while interacting with a discriminatory object were analyzed. Most of the interactors understood the interaction to be done with the unethical machine and the meaning of this metaphor. Some interactors demonstrated unsatisfied with the ethnicity results that led to make user to try again those results, to contrast with other interactors and in some

cases, a reflection on errors that AI produce and their dataset. Moreover, there were noted some behaviors with the use of the device and the impact of receiving a benefit (gumball) or not. Few cases, interactors tried to modify the results by positioning an interactor that already was recognized as white to unlock the machine and share gumballs with the others or even in one case, using an image of a white male in a smartphone to unlock the machine. Other interactors that were recognized as white decided to not to get the gumball or share it with others. There was also reported a few discussions generated after the use of *BOX* around our main topic: ethnicity in artificial intelligence. After these preliminary results, we aim to continue creating more unethical machines and explore a deeper evaluation on the study field with the use of interviews and questionnaires [18].



Figure 3. *BOX*'s interaction.

Final Discussion

Unethical Machines aims to explore the social impact of AI introduced in everyday life. This exploration has a dual focus: we are interested in design opportunities that may arise, as well as in the ethical concerns that the popularization of this technology entails.

The project is designed to allow the general public to experience in person the power of machine learning, computer vision, and natural language processing in a familiar environment, while simultaneously providing a tool for participatory observation, as well as ethnographic and technographic research.

Technographic research is often missing in AI studies. However, technography's focus on what technology actually does, allows researchers to map the concrete behavior of automated objects and the sets of beliefs that social groups hold. With this in mind, we are implementing participatory action research, documenting interactors' behaviors with the device.

The project's future work includes developing new mundane objects that will also adjust their behavior in function of predictions of ethically contentious characteristics, such as sexual orientation, ethnicity, age, income, etc. Future work will also include representatives from communities that are often disenfranchised from the technology produc-

tion process, as well as from communities at risk of being negatively affected by AI technologies.

Acknowledgements

The authors thank Stochastic Labs for supporting the development of the project. The work described in this paper partially supported by Tomas Laurenzo's grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. 9042456 (CityU 11674416)]. Katia Vega also thank the support grant from Johnson and Johnson WiSTEM2D Scholars Award, and CoR Small Grant from UC Davis Academic Senate.

References

- [1] Affectiva. 2018. Affectiva. <https://www.affectiva.com/how/how-it-works/>. (2018).
- [2] Vicki Bruce and Andy Young. 1986. Understanding face recognition. *British journal of psychology* 77, 3 (1986), 305–327.
- [3] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
- [4] Sheelagh Carpendale. 2008. Evaluating information visualizations. In *Information visualization*. Springer, 19–45.
- [5] Clarifai. 2018. Clarifai Demographics Model. <https://clarifai.com/models/>. (2018).
- [6] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.
- [7] Haystack.ai. 2018. Haystack. <https://www.haystack.ai/>. (2018).
- [8] Tomas Laurenzo and Katia Vega. 2019a. DOOR: Ethnicity in Artificial Intelligence. In *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 535–540.
- [9] Tom Laurenzo and Katia Vega. 2019b. Ethnicity in Artificial Intelligence. *LatinX in AI Research at ICML 2019*. (June 2019). <https://hal.archives-ouvertes.fr/hal-02265913> Poster.
- [10] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.
- [11] Ann J Morrison, Peta Mitchell, and Margot Brereton. 2007. The lens of ludic engagement: evaluating participation in interactive art installations. In *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 509–512.
- [12] Future of Privacy Forum. 2017. Unfairness y Algorithm: Distilling the Harms of Automated Decision-Making. <https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf>. (2017).
- [13] Ashok Samal and Prasana A Iyengar. 1992. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern recognition* 25, 1 (1992), 65–77.

- [14] David Sandberg. 2017. Face recognition using tensorflow. Läst [2019-11-20]. URL: <https://github.com/davidsandberg/facenet> (2017).
- [15] New Scientist. 2018. How Google and Facebook hooked us and how to break the habit. <https://www.newscientist.com/article/mg23731640-500-how-google-and-facebook-hooked-us-and-how-to-break-the-habit/>. (2018).
- [16] Song Yang Zhang, Zhifei and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
- [17] TEI Arts Track. 2019. BOX. Ethnicity in Artificial Intelligence. Last [2019-12-16]. URL: <https://tei.acm.org/2019/program/arts-track.html>
- [18] Candy L, Amitani S, Bilda Z. Practice-led strategies for interactive art research. CoDesign. 2006 Dec 1;2(4):209-23.

Authors Biographies

Tomas Laurenzo is an Assistant Professor at the School of Creative Media of the City University of Hong Kong. He worked as associate professor at University of the Republic (Uruguay), where he founded and directed both the Medialab of the Engineering School and the university's Core Group on Human-Computer Interaction. He also worked as Associate Researcher at the Centre for Basic Research on Psychology, Visiting Professor at the School of Architecture and Design, and was a fellow of both the National Agency of Research (ANII) and the PEDECIBA program, the main research-funding institutions in Uruguay. Laurenzo has also performed artistic and academic activities in several institutions including The Broad Institute of Harvard and MIT, Microsoft Research, Carnegie Mellon University, Brunel University, Ericsson, University of Iowa, Stochastic Labs, and INRIA, among others. He holds a Ph.D. in Computer Science from the PEDECIBA Program, advised by Dr. Alvaro Cassinelli, University of Tokyo.

Katia Vega is an Assistant Professor at the Department of Design in the University of California, Davis. She was a Postdoc Associate at MIT Media Lab (USA). She got her PhD and Master Degree in Computer Science at PUC-Rio (Brazil). Her research interests include wearable technologies, interactive tattoos, beauty technology and skin interfaces. Her work has been featured by New Scientist, Wired, Discovery, CNN, among others. In 2016, MIT Technology Review awarded her as one of the 5 Innovators under 35 in Peru. Recently, CNET recognized her as one of the Top 20 most influential Latinos in Tech in 2017. SXSW 2018 awarded The Dermal Abyss project in the Science Fiction no Longer Category. In 2019, she was honored as a leading woman in STEM by Johnson & Johnson..



ISEA International

Copyright License Agreement

Archives and Promotion Form

This copyright release form is to be used by ISEA International for permission to use materials for the ISEA Archives and promotion of ISEA in general. The materials in question include, but are not limited to, papers, abstracts, biographies, proposals, artworks, videos and animations, documentation of performances, and other creative events or visual, textual, or auditory artifacts, hereinafter collectively known as the *Works*. If the *Works* were produced by multiple authors/creators, each must submit a separate Copyright License Agreement form.

By signing this form I hereby grant ISEA International revocable, non-exclusive, perpetual, royalty free, unrestricted worldwide rights to publish and distribute the *Works* for scholarly use in the ISEA Archives and for promotional purposes. The *Works* may be reproduced, cropped, or edited for grammatical error or clarity and presented in any media now known or hereafter to become known, including, but not limited to, print and all formats of electronic media.

ISEA International agrees to credit the Copyright Owner wherever it is possible and space permits and will respect the integrity of the *Works*.

I certify that the *Works* are original artifacts and I, along with collaborators if indicated below, am the sole owner(s) and have the requisite rights and permissions to make this agreement. If the *Works* contain material created by a third party, I have obtained any and all required permissions in writing from all third party copyright holders for the reproduction of any included images, audio, movies, and text.

I indemnify *ISEA* against any claim, loss, liability, cost or expense that may be incurred by *ISEA* as a result of any breach of this contract.

Contributor's Information:

Signature: _____

Date: 12-16-19

Name: Katia Vega

Title of *Works*: Exploring AI Ethics Through Intelligent Everyday Objects

Affiliation: UC Davis

Description/Type: Short Paper

Address: One Shields Ave, Davis, CA.

Email: kvega@ucdavis.edu

Credit/Copyright line/s to be used:

Collaborators: Katia Vega, Tomas Laurenzo

2019, Tomas Laurenzo, Katia Vega.

Copyright License Agreement

Artist/Performer Form

This copyright release form is to be used for artworks, videos and animations, documentation of performances (including but not limited to musical, theatrical, artistic, and dance), and other creative events or visual and audio artifacts, hereinafter collectively known as the *Works*. If the *Works* were produced by multiple authors/creators, each must submit a separate Copyright License Agreement form.

By signing this form I hereby grant ISEA2020, ISEA International, its affiliates, partners, and those acting with its authority (hereinafter collectively referred to as "ISEA"), revocable, non-exclusive, perpetual, royalty free, unrestricted worldwide rights to exhibit the *Works* in an ISEA exhibition, publish the *Works* in ISEA documents, transmit the *Works* via electronic media, and distribute the *Works* in ISEA materials for scholarly activities, promotion, dissemination of information, preservation in the ISEA archives, and documentation of ISEA events. The *Works* may be reproduced, copied, cropped, or modified for clarity and presented in any media now known or hereafter to become known, including, but not limited to, print and all formats of electronic media.

ISEA agrees to credit the Copyright Owner wherever it is possible and will respect the integrity of the *Works*.

I certify that the *Works* are original artifacts and I, along with collaborators if indicated below, am the sole owner(s) and have the requisite rights and permissions to make this agreement. If the *Works* contain material created by a third party, I have obtained any and all required permissions in writing from all third party copyright holders for the reproduction of any included images, audio, movies, and text.

I indemnify ISEA against any claim, loss, liability, cost or expense that may be incurred by ISEA as a result of any breach of this contract.

Contributor's Information:

Signature: _____



Date: 12-16-19

Name: Katia Vega

Title of *Works*: Exploring AI Ethics Through Intelligent Everyday Objects

Affiliation: UC Davis

Description/Type: _____

Address: One Shields Ave, Davis, CA.

Email: kvega@ucdavis.edu

Credit/Copyright line/s to be used: _____

Collaborators: Katia Vega, Tomas Laurenzo

2019, Tomas Laurenzo, Katia Vega.

Copyright License Agreement

Presenter/Author Form

This copyright release form is to be used by presenters, authors and workshop facilitators. The artifacts produced may include, but is not limited to, abstracts, academic papers, proposals, images, videos, audio recordings, hereinafter collectively will be known as the *Works*. If the *Works* were produced by multiple authors, each must submit a separate Copyright License Agreement form.

By signing this form I/we hereby grant ISEA2020, ISEA International, its affiliates, partners, and those acting with its authority (hereinafter collectively referred to as "*ISEA*"), *revocable*, non-exclusive, perpetual, royalty free, unrestricted worldwide rights to present, publish, transmit, and distribute the *Works* for scholarly activities, promotion, dissemination of information, preservation in the ISEA archives, and documentation of *ISEA* events. The *Works* may be reproduced, translated, or edited for purposes of clarity, format, and grammar and presented in any media now known or hereafter to become known, including, but not limited to, print and all formats of electronic media.

The *Works* may be, but is not guaranteed to be, published electronically or in hard copy in the proceedings, book of abstracts, catalog, archives, host website, or other symposium artifacts. *ISEA* agrees to credit the Copyright Owner wherever it is customary and space permits and respect the integrity of the *Works*.

A paper submitted to *ISEA* should not have been published previously, nor should it be submitted to another publisher or symposium at the same time. If the author re-publishes the paper in any form after acceptance by *ISEA*, they will acknowledge *ISEA* as the original publisher.

I certify that the *Works* are original creations and I, along with collaborators if indicated below, am the sole owner(s) and have the requisite rights and permissions to make this agreement. If the *Works* contain material created by a third party, I have obtained any and all required permissions in writing from all third party copyright holders for the reproduction of any included illustrations, photos, audio, movies, and extended quotations.

I indemnify *ISEA* against any claim, loss, liability, cost or expense that may be incurred by *ISEA* as a result of any breach of this contract.

Contributor's Information:

Signature: _____



Date: 12-16-19

Name: Katia Vega

Title of *Works*: Exploring AI Ethics Through Intelligent Everyday Objects

Affiliation: UC Davis

Description/Type: Short Paper

Address: One Shields Ave, Davis, CA.

Credit/Copyright line/s to be used:
2019, Tomas Laurenzo, Katia Vega.

Email: kvega@ucdavis.edu

Collaborators: Katia Vega, Tomas Laurenzo

Comments/Description: _____



ISEA2020 and ISEA International **Copyright License Agreement**

Consent to be Audio/Video Recorded

This copyright release form is to be used for audio or video recording of speakers, presenters, or participants. If an event includes multiple participants who are recorded, each must submit a separate Copyright License Agreement form.

I hereby give my consent to ISEA2020, ISEA International, its affiliates, partners, and those acting with its authority (hereinafter collectively referred to as "ISEA"), to record my participation in the program or event described below (hereinafter, the *Recording*).

I understand and agree that the content of the presentation as well as my name, likeness, image, and/or voice may be recorded and the *Recording* will be retained and archived by ISEA.

I understand and agree that ISEA holds complete ownership and copyright of the *Recording* in which I or my presentation, performance or contribution appears.

I expressly release ISEA, and all persons acting under its permission or authority, from every claim, liability, or action in law, including any and all claims for defamation or copyright infringement, arising as a result of the production and distribution of the *Recording* of my participation in this event.

Use & distribution of the recording:

I hereby grant a non-exclusive, royalty-free license to ISEA to retain, reproduce, and distribute worldwide this *Recording* for non-commercial purposes. I understand and agree that this *Recording* may, but is not guaranteed to be included in ISEA's Archives and other online ISEA resources.

This *Recording* may be recorded on or transferred to analog or digital media in formats now known or hereafter developed. It may be used, distributed, edited, mixed, and duplicated without restriction, in whole or in part, by ISEA for publication or transmission via means now known or hereafter developed. It may be used for any scholarly and promotional use deemed appropriate by ISEA.

Contributor's Information:

Signature: _____

A handwritten signature in blue ink, appearing to read "Katia Vega", written over a horizontal line.

Date: 12-16-19

Name: Katia Vega

Title of Works: Exploring AI Ethics Through Intelligent Everyday Objects

Affiliation: UC Davis

Description/Type: Short paper

Address: One Shields Ave, Davis, CA.

Website: katiavega.com

Email: kvega@ucdavis.edu

