Policy and ML-Assisted Decision Making:

The Need for Multi-Resolution, Multi-Strength, Interactive Explanations of Learned Systems

David Bayani *
dcbayani@andrew.cmu.edu
Stefan Mitsch
smitsch@cs.cmu.edu
Carnegie Mellon University
School of Computer Science
Pittsburgh, Pennsylvania

April 23, 2020

* Speaking

This material is based upon work supported by the United States Air Force and DARPA under Contract No. FA8750-18-C-0092. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force and DARPA.

2020 CMU Symposium on AI and Social Good

Broad Involvement of AI in Current and Future Policy Making

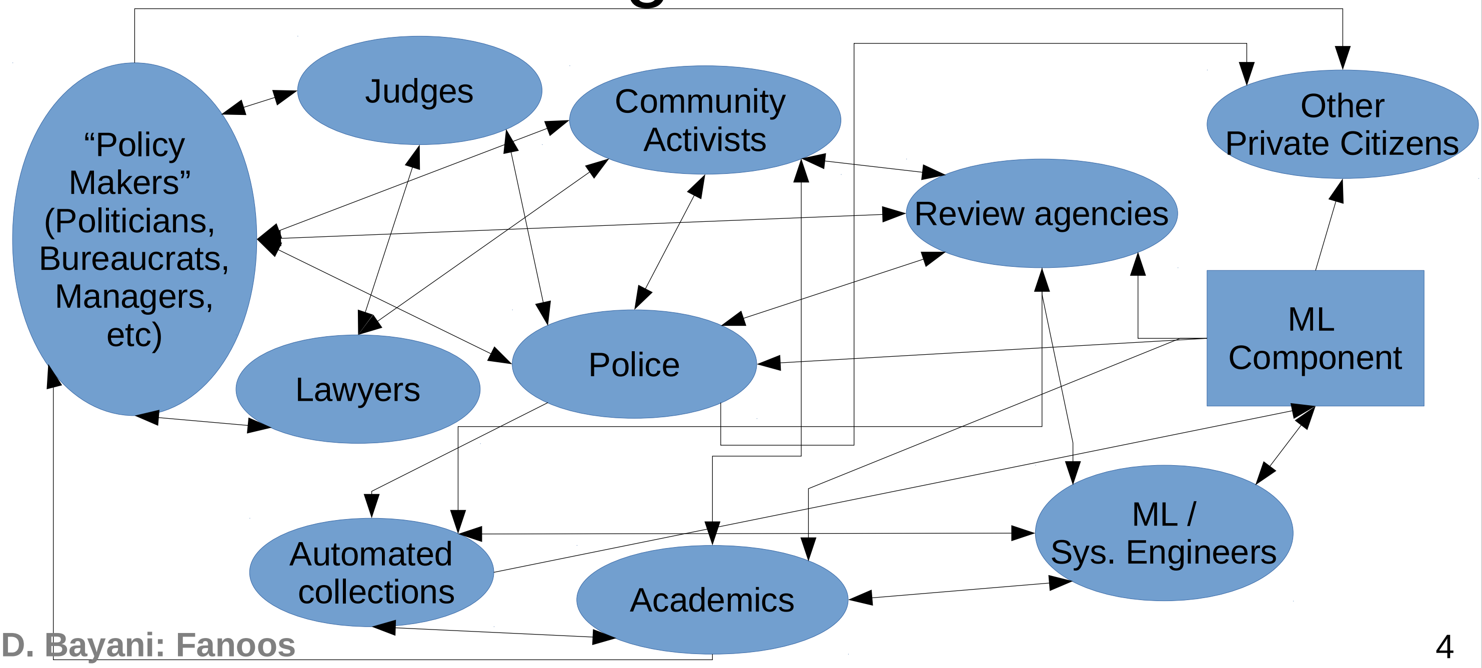
- Attempted applications span most areas of life
  - Primary focus here: ***Machine Learning (ML)-based*** AI

Broad Involvement of AI in Current and Future Policy Making

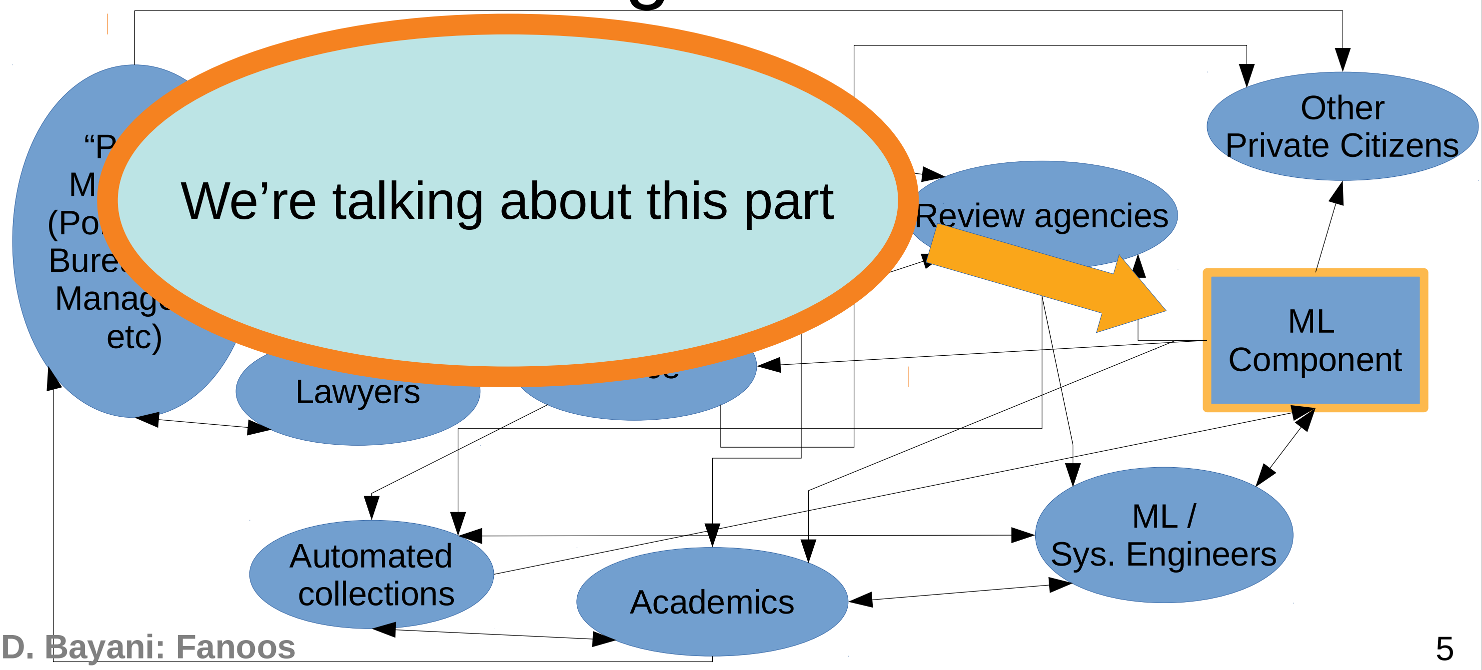
- Attempted applications span most areas of life
  - Primary focus here: ***Machine Learning (ML)-based*** AI
- Example with clear direct AI-Impact: ***Law-and-Order***
  - E.g., Recidivism prediction
  - Concerns about ***fairness*** (e.g., [10,11])

Examples for Law-and-order:Recidivism prediction, predictive policing, detecting fake government contracts

AI is part of a (much) larger whole in tackling social issues



AI is part of a (much) larger whole in tackling social issues



Some Observations

- High stakes: want to check what ML component does
  - Explainable AI : ***XAI***
- Actors have many different:
  - Interests
  - Needs
  - Backgrounds / Expertise

All the actors are invested in the outcome of the process. If the ML component holds any actual sway on the outcomes, many may want to check what it s doing or has learned. Ethical, legal, community, and public-relations impacts can all be at stake. These concerns motivate desires for Explainable AI (XAI) in this application area. However, given the variety of backgrounds, needs, roles, and interests of different actors, one type of explanation likely will not serve many of those involved in the process well. Explanations need be curated to their audience to improve interpretability.

Current State and Trends in XAI

- Fast growing ([1]), but currently has limitations:
  - Explanations of *single* granularity/abstraction level
  - Lack of guarentees
  - Interaction is usually to explore data, not models

Conclusions: Unfilled Desiderata for XAI in Social Good

- *Interactive* (so can explore as needed)

Conclusions: Unfilled Desiderata for XAI in Social Good

- *Interactive* (so can explore as needed)
- *Can provide multiple abstraction levels of information* (so can suite multiple audiences and needs)

Conclusions: Unfilled Desiderata for XAI in Social Good

- *Interactive* (so can explore as needed)
- *Can provide multiple abstraction levels of information* (so can suite multiple audiences and needs)
- Can provide strong guarentees about info. provided (so *explanations necessarily reflect system behaviour*)
  - Should be as pedantic about details as user needs (*sometimes want / don't want corner-case info.*)

Our Solution: Fanoos

- Fanoos (فــانوس) - “Lantern” in Farsi

“Shining a Light on Black-Box AI”

Masrawy.com

This system is our attempt to meet these desiderata. One of our hopes in presenting this talk is to inspire others to go down a similar path. There is a lot of room for expansion and improve, and the utility of having such systems is clear.

Fanoos Overview:

Initial Setup

Fanoos

Fanoos Overview:

Initial Setup

L, the Learned Sytem to Explain (E.g., Neural Net)

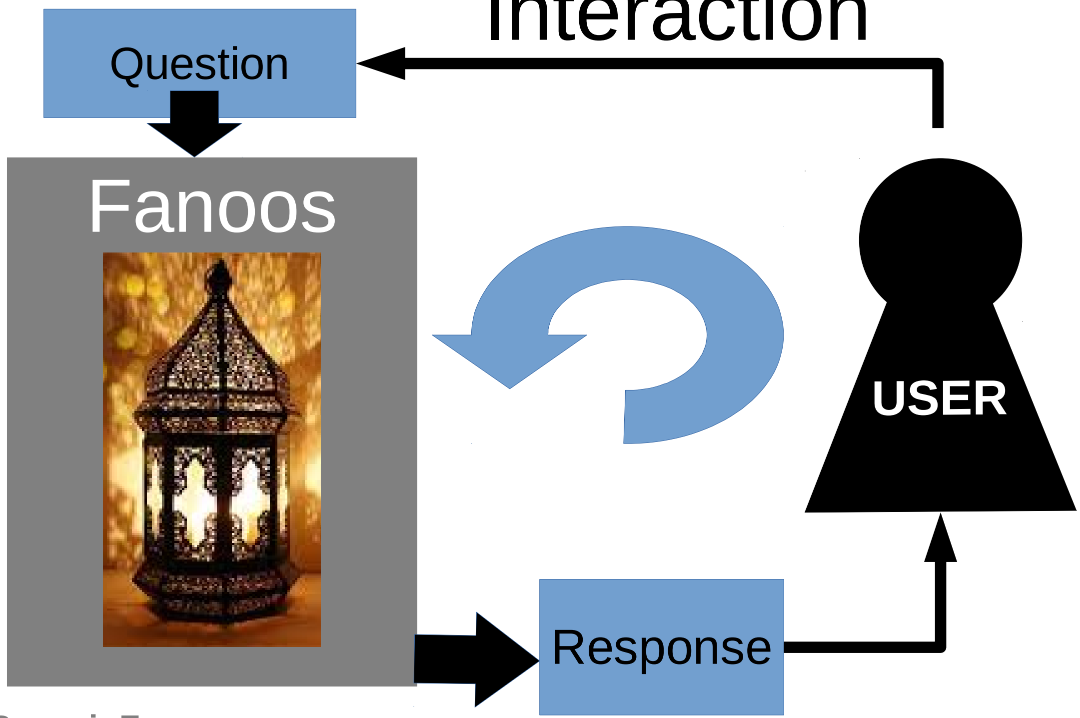
Domain Info. (can have data-assisted construction)

Initial Configuration Inputs

Fanoos

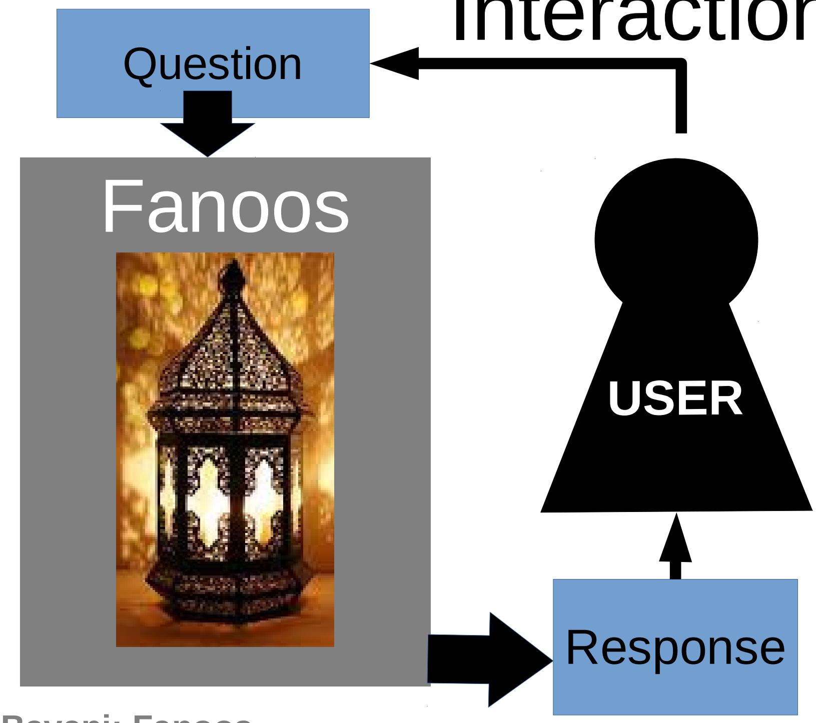
Fanoos Overview:

Interaction



Fanoos Overview:

Interaction



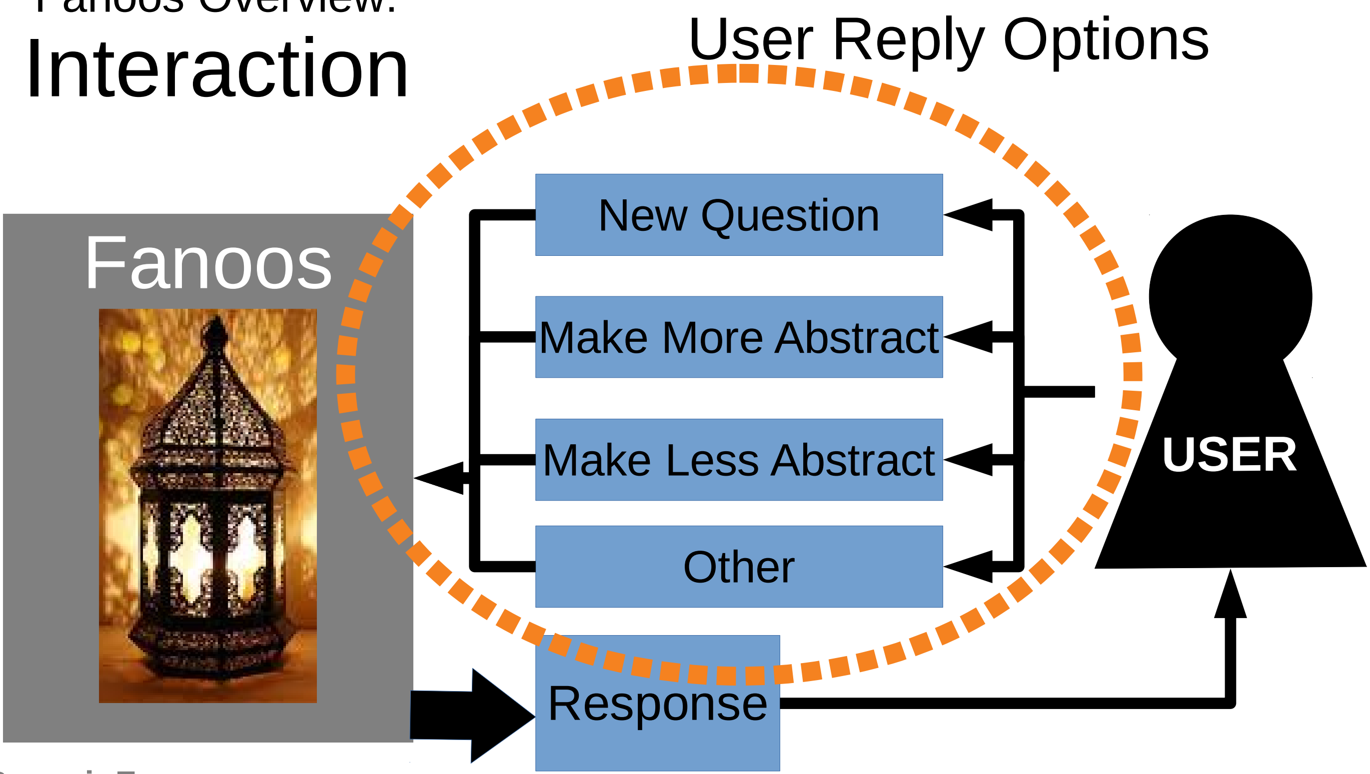
Question Types:

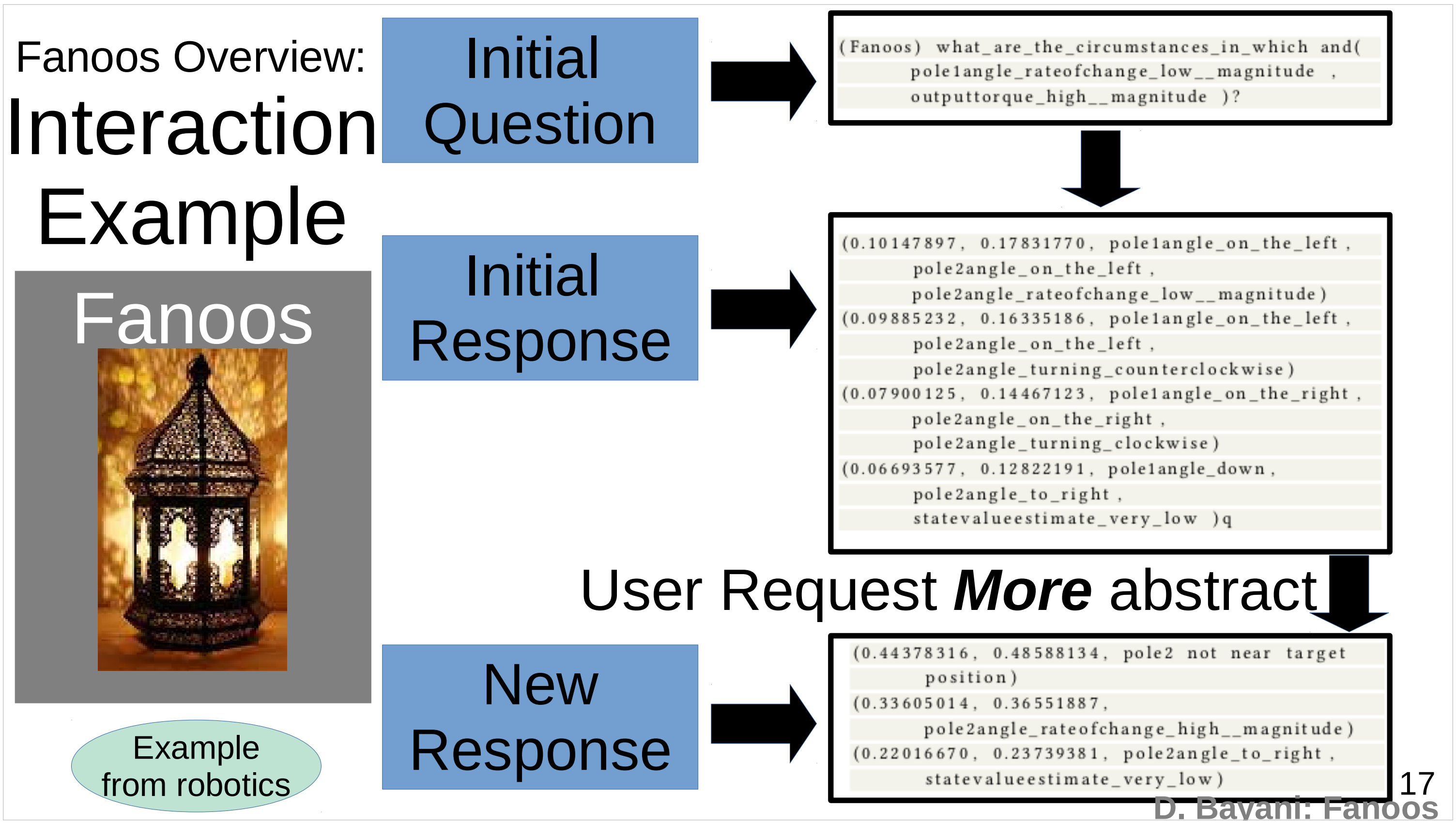
- When does ***L*** do X?
- What does ***L*** do when Y?
- In what circumstances is ***L*** doing X during Y?

Can be formally sound or probabilistically guaranteed

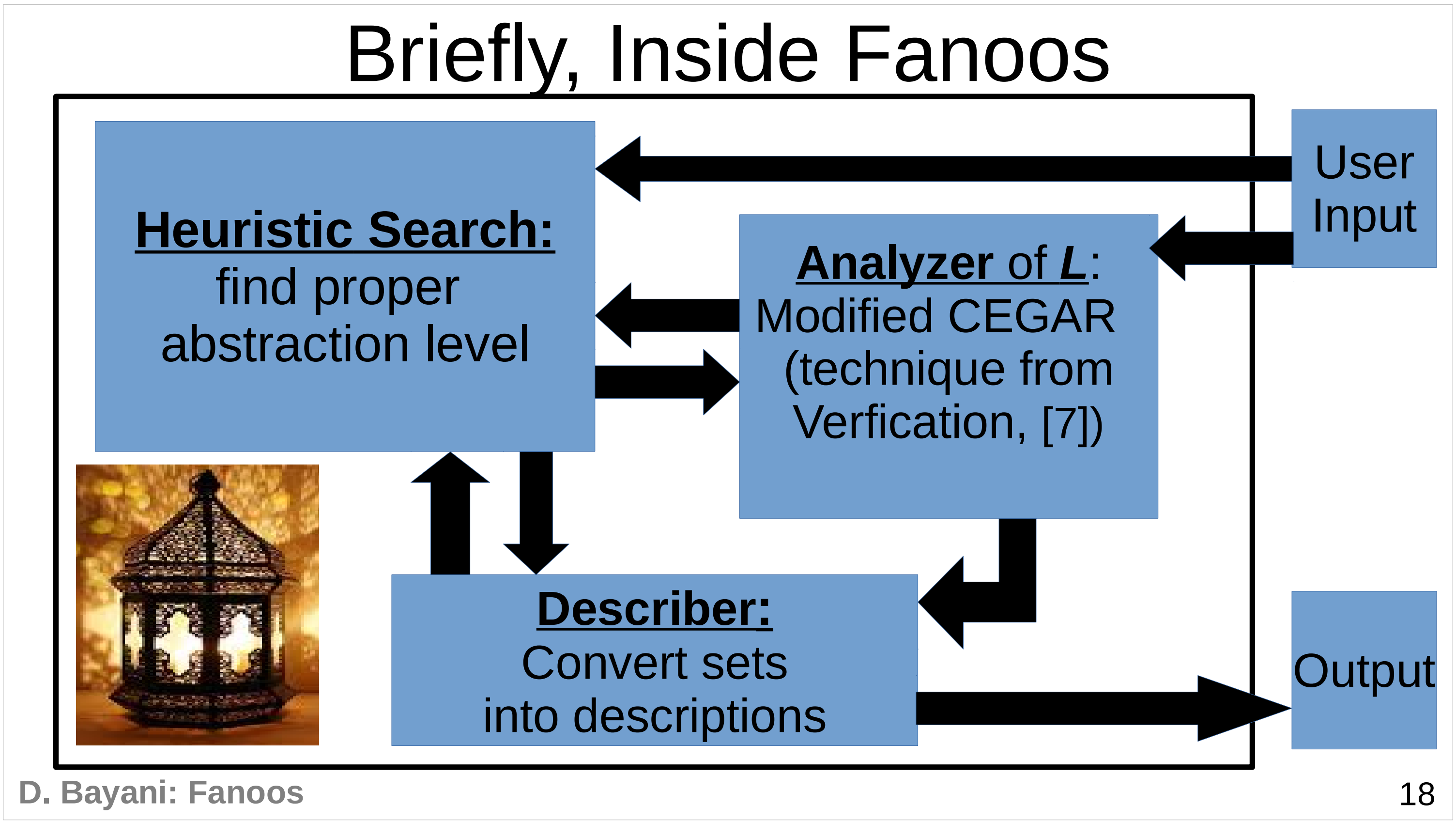
Fanoos Overview:

Interaction





The new response is both shorter and contains different verbiage.



This schematic is a simplification of the components and their interactions in Fanoos. While not displaying everything, it does convey the main spirit of information flows in our approach.

Heuristic Search: Key idea: view finding the proper description abstraction as an informed search over system states and parameters

This is an important part to the design philosophy of Fanoos. As an analogy, for instance, one can think of user feedback as informing what branch to take in a binary search for the proper description state. Fanoos involves more sophisticated machinery and search spaces than simply this, but the spirit is similar.

Analyzer of L: Based on modified CEGAR algorithm- a classic verification technique

See the references for some pointers to CEGAR.

Describer: Convert sets discovered into reasonable description

Fanoos example uses:

- ***Pre-deployment*** detection of unfair treatment toward people with ***joint-protected attributes***
  - E.g., Gender G and Race R as separate wholes may not be disparately treated, but those with G-R attributes may be.(see [8])
- **counter-factual assessment** with guarantees
- Coming application to neural-net trained to output credit-scores (in country outside USA) (see [2])

Be aware: XAI is not a Cure-all

- XAI is not a panacea for fairness concerns ([3])
- Need domain experts in the loop
- Fanoos tries to improve the situation: suit more actors

General Take-Aways and Thoughts for Future Work

- Need for flexibility and ***varying abstraction***
  - Examples of this working well for other tools – home computers (GUI, advance settings, terminals, Python, assembly, hardware)
- Under-explored: ***Formal Verification + XAI***
  - A lot of work open to explore
  - Prior work in XAI for social applications largely either:
    - does post-facto analysis (e.g., [14,15], sec. 10 in [13])
    - use basic models (e.g., [4])

References and Further Reading

Further details can be found at the author's github page at a later date:
https://github.com/DBay-ani

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Philip Adler, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing Black-box Models for Indirect Influence. (2016). arXiv:stat.ML/1602.07043

[3] Solon Barocas. 2014. Data mining and the discourse on discrimination. In *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining*. 1–4.

[4] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[5] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.

[6] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056* (2017).

[7] Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, and Helmut Veith. 2000. Counterexample-guided abstraction refinement. In *International Conference on Computer Aided Verification*. Springer, 154–169.

[8] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[9] Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (2017), 14–29.

[10] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).

[11] John Lightbourne. 2017. Damned Lies & Criminal Sentencing Using Evidence-Based Tools. *Duke Law & Technology Review* 15, 1 (2017), 327–343.

[12] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.

[13] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 29, 5 (2014), 582–638.

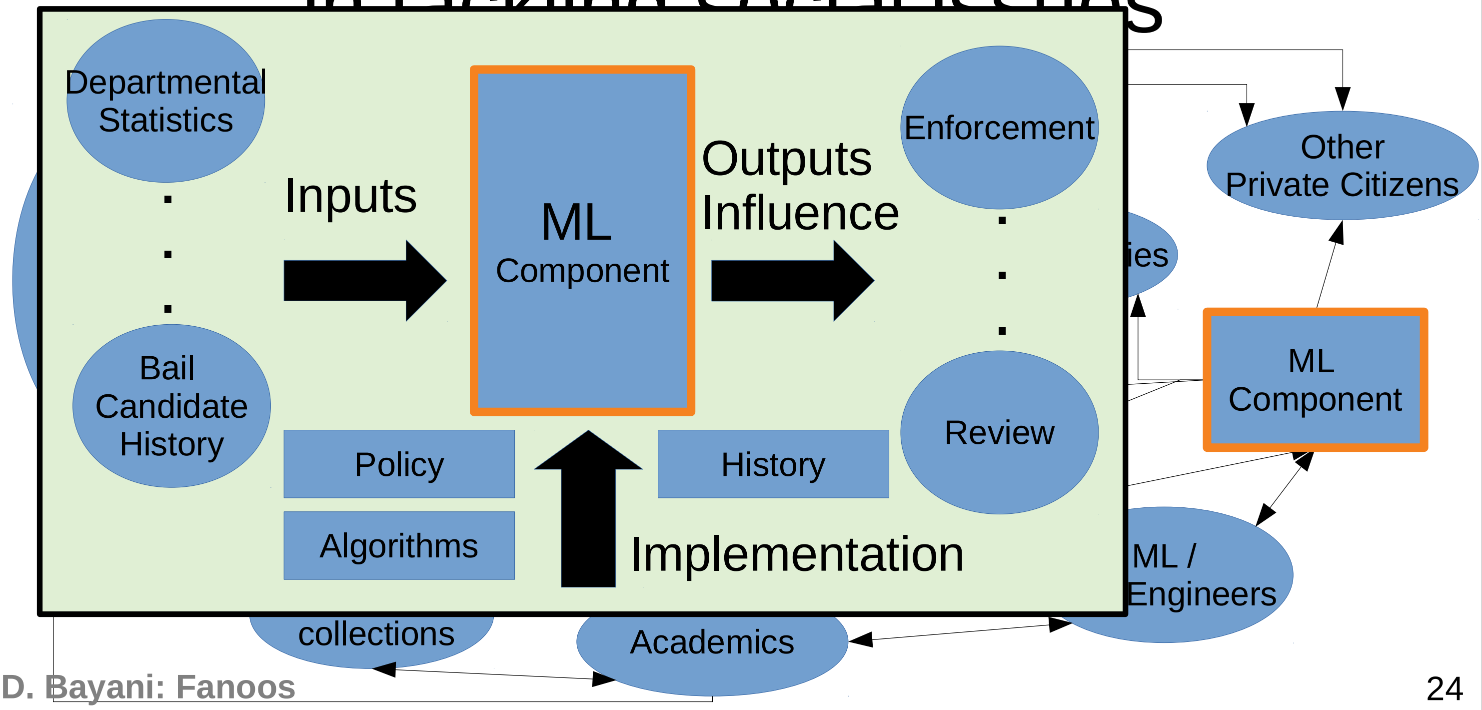
[14] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[15] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014).

[16] Tal Zarsky. 2013. Transparent Predictions. *University of Illinois Law Review* 2013, 4 (2013).

Appendix

AI is part of a (much) larger whole in tackling social issues



Be aware: XAI is not a Cure-all

- XAI is not a panacea for fairness concerns([3])
  - Context is key
  - Garbage-in-garbage-out principle ([3])
- Need domain experts in the loop. Examples why:
  - Simpson's Paradox
  - proxies for protected categories (see [13,16])
  - A "fair" model might produce "unfair results" in the world, and vice-versa (see [9])
    - The input distribution from world can produce "unfair" output distribution, even if function imbetween is even-handed
- Fanoos tries to improve the situation
  - more integration by those with the context-knowledge
  - May allow for discovery of previously unknown proxy variables and confounders

Some General Take-Aways and Thoughts for Future Work

- Need for abstraction and flexibility
  - Likely would benefit from standardization of “abstraction levels”, particularly at divisions suitable for policy makers
  - Examples of abstraction levels working well for other tools – home computers (GUI, advance settings, terminals, Python, C, assembly, hardware)
- Under-explored in recent years: formal methods from program analysis/verification leveraged for interpretable ML (formal verification + XAI)
  - A lot of complementary abilities, focuses and available exploration here
  - Prior work in XAI for social applications is largely post-facto analysis (e.g., [14,15], sec. 10 in [13]) or deals with manual examination of simple models (e.g., [4])
- Probabilistic and formal guarantees complement each other
  - Can capture different properties (e.g. include/ignore measure-zero sets)
  - Neither dominates – different situations call for different choice
- Many tools have pros and cons: try to make it easy to use the right ones at the right times
  - E.g.: disparate-impact assessment versus model-audit
  - Also may help support more nuanced evaluations. e.g.: evaluation across different fairness metrics (e.g. [6,14]).