Focus on values: justice, fairness, and related concepts

P. J. Nickel

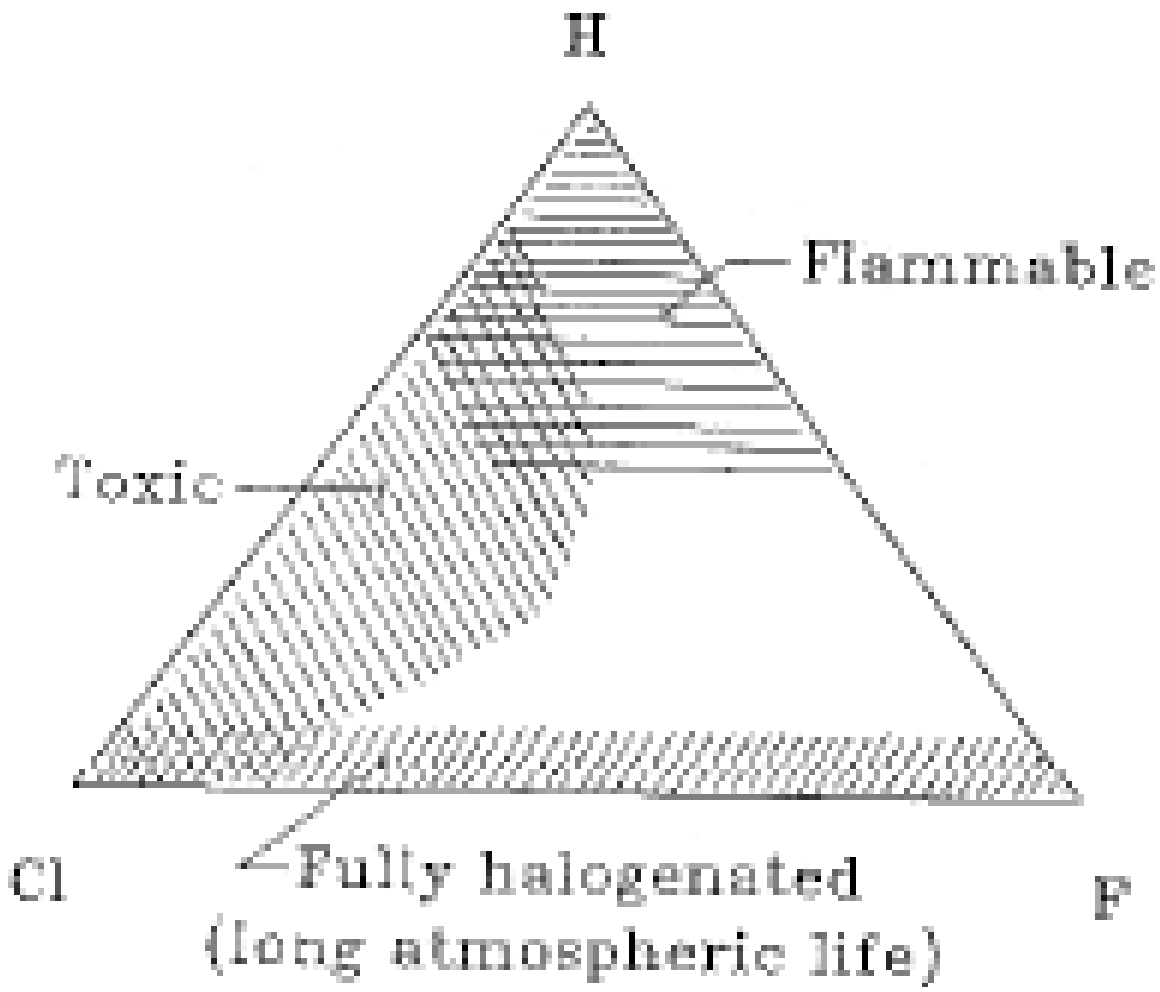
Associate Professor, Dept of Philosophy & Ethics

Agenda

- Review “Design for Values”
- Justice as (algorithmic) fairness
- Example of digital health
- Relational justice
- Punitive justice

Review “Design for Values”

diagram of ethical tradeoffs in refrigerants



Ibo van de Poel, Investigating Ethical Issues in Engineering Design, *Science and Engineering Ethics* 7, 3 (2001): 429-446.

Incorporating ethics into AI applications: some approaches

1. Build the system using ethical rules/ values.
("Whenever you detect a human, stop the robot" [SAFETY]
"Do not make more systematic errors for any sub-group in the dataset compared to the others" [FAIRNESS])
2. Train systems on large data sets in which the rules are followed (e.g., cases where the robot stops for humans)
3. Train systems on large data sets in which the rules might not have been followed, but impose the rules as filters.
4. Rely on the socio-technical system (users) to provide ethical input.

Cases from CHLEP Ch. 3

Generative language models
AI for determining loan eligibility
AI for benefits fraud detection

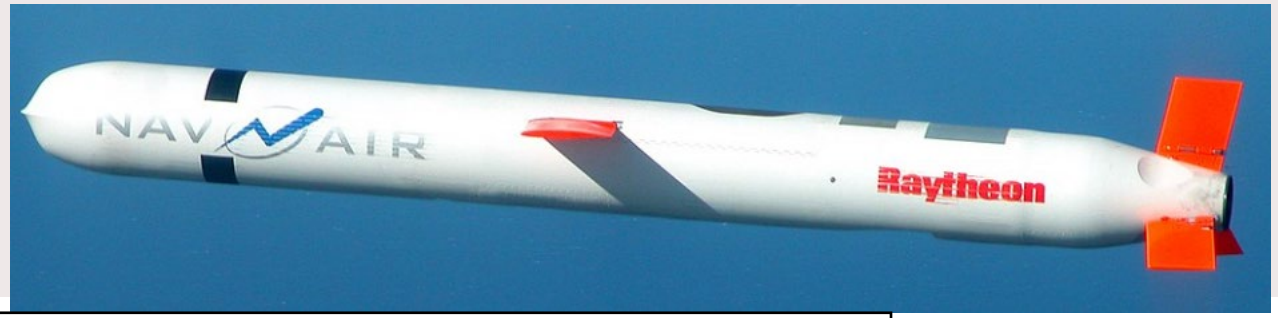
Please look carefully at how the authors describe these cases ☺

Some limitations of Design for Values

- Some ethical issues appear too late in the design process to be incorporated into the design.
- Design for Values is very resource-intensive, especially if it has to be specified for each application.
- In cases of high moral uncertainty (‘moral disruption’), it may be impossible to specify values in the way required by Design for Values.

Some limitations of Design for Values

- Some ethical issues appear too late in the design process to be incorporated into the design.
- Design for Values is very resource-intensive, especially if it has to be specified for each application.
- In cases of high moral uncertainty (‘moral disruption’), it may be impossible to specify values in the way required by Design for Values.
- Some technologies may be impossible to ethically “optimize”



https://commons.wikimedia.org/wiki/File:Tomahawk_Block_IV_cruise_missile_-crop.jpgmedia Commons image source

Military ethics and VSD: an example of where values come from in DfV/ VSD

<u>just war theory</u>

Guiding assumption: The only justification for war is self-defense.

Evidence that nations takes this seriously in principle: Department of War → Department of Defense

jus ad bellum* and *jus in bello

principles of proportionality and discrimination

the role of ethics?

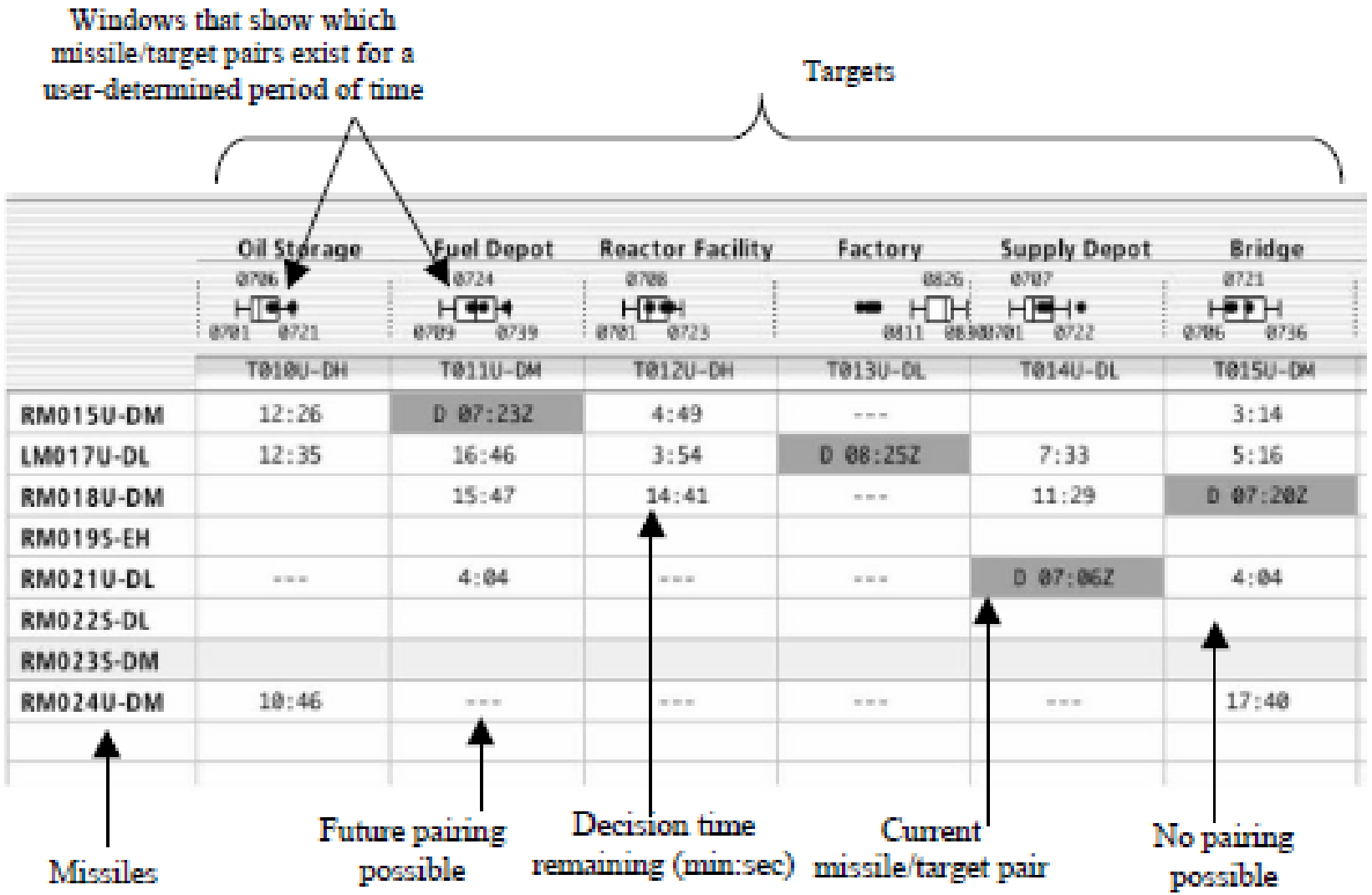
“The primary purpose of analyzing conceptual ethical issues in the context of a military command and control system is to identify human value design considerations that will guide the engineer in constructing a more effective decision support system. For example, understanding the principle of discrimination can guide the engineer in developing a system that accounts for the risk of harm to both friendly and non-combatant forces” (Cummings 2006, 706).

the tactical Tomahawk

"The Tomahawk missile is the U.S. Navy's premier land attack missile ... once launched, [it] cannot be redirected in-flight. This limitation to the system causes potential redundant demolition of targets, thus wasting the $1.2 million dollar [*sic*] missile, as well as the inability to correct for a targeting mistake in flight. In addition ... the missile cannot respond to dynamic situations in which, for example, a target has moved, or a more critical target emerges."

→ develop Tactical Tomahawk with re-direction ability

decision-support tool



level of automation

Automation Level	Automation Description
1	The computer offers no assistance: human must take all decisions and actions.
2	The computer offers a complete set of decision/action alternatives, or
3	narrows the selection down to a few, or
4	suggests one alternative, and
5	executes that suggestion if the human approves, or
6	allows the human a restricted time to veto before automatic execution, or
7	executes automatically, then necessarily informs humans, and
8	informs the human only if asked, or
9	informs the human only if it, the computer, decides to.
10	The computer decides everything and acts autonomously, ignoring the human.

In the empirical phase, what emerged? ---Automation bias

- Human operators have a tendency to accept the default of automated, “smart” systems. When the selection system for the missile highlighted a target, operators tended to accept that target, even if they possessed other information that should have led them to double-check it.
- This does not directly test well-being, but it does suggest a problem for the principle of human autonomy, and maybe for responsibility.

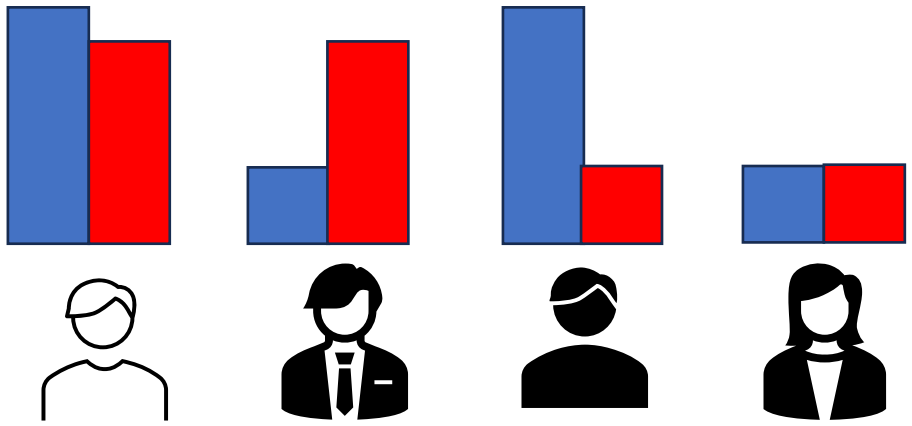
Justice as (algorithmic) fairness

Algorithmic fairness must represent relevant levels of well-being or need for different individuals

1. Suppose a given application is algorithmically fair.
2. If fairness is algorithmic, the elements that make a distribution fair must not be accidental. (Follows from robustness.)
3. For the elements that make a distribution fair to be non-accidental, they must be represented in the algorithm.
4. The elements that make a distribution fair are an individual's prior well-being or need, as well as the change to their condition that is predicted; as well as a representation of parity between groups.
5. Therefore, the application must represent all relevant individuals' prior well being or need, the predicted change to their condition, and group membership.

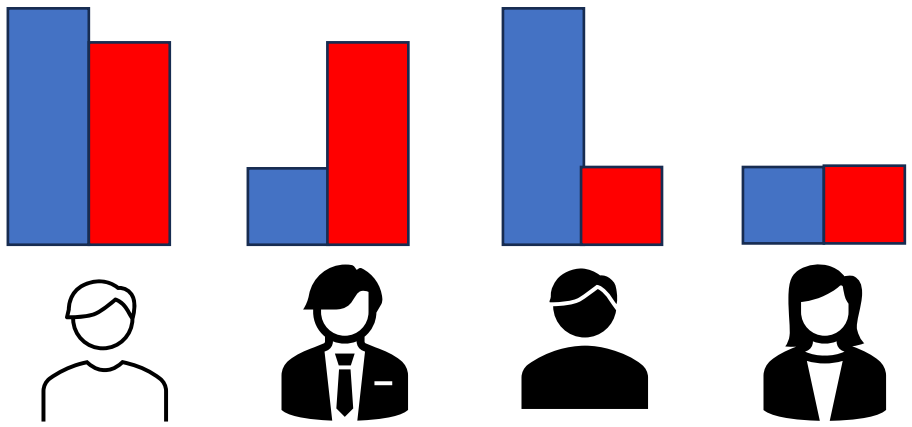
Distributions of benefit

- Current well-being or need = red
- Amount of benefit predicted = blue



Distributions of benefit

- Current well-being or need = red
- Amount of benefit predicted = blue

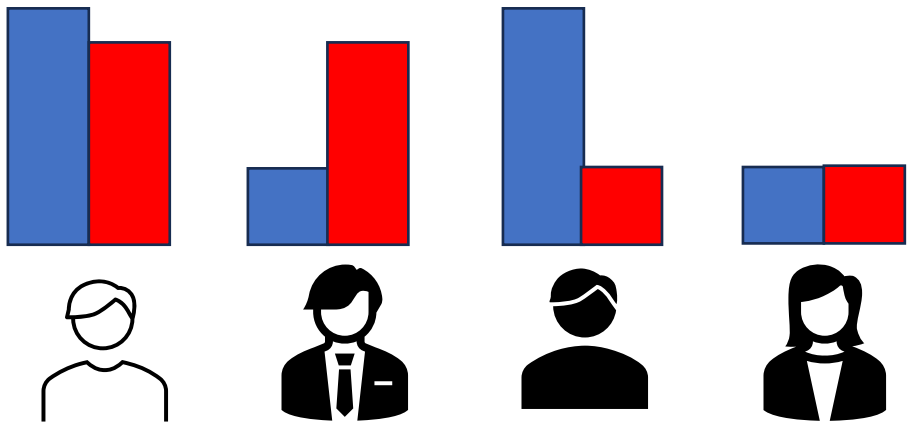


Fairness through intervention. This approach aims to attain algorithmic fairness by directly introducing technical algorithmic interventions either during the model development phase (i.e. in-processing[1]) or afterwards as an additional layer (i.e. post-processing[2]). Within a model development pipeline, in order to detect and mitigate bias, alternative fairness interventions techniques are available and used in practice. Most commonly, to the best of our knowledge, fairness interventions are implemented by means of a *Parity-based approach*, aiming to achieve a form of parity of outcomes between different groups. Below we describe the main idea behind

Barsotti, Flavia, and Rüya Gökhan Koçer. "MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias." *AI & SOCIETY* 39, no. 3 (2024): 961-974.

Distributions of benefit

- Current well-being or need = red
- Amount of benefit predicted = blue



According to the context and our theoretical commitments, we can choose metrics for well-being/ need and benefit such as:

Wealth
Capability
Health

Algorithmic fairness must represent relevant levels of well-being or need for different individuals

1. Suppose a given application is algorithmically fair.
2. If fairness is algorithmic, the elements that make a distribution fair must not be accidental. (Follows from robustness.)
3. For the elements that make a distribution fair to be non-accidental, they must be represented in the algorithm.
4. The elements that make a distribution fair are an individual's prior well-being or need, as well as the change to their condition that is predicted; as well as a representation of parity between groups.
5. Therefore, the application must represent all relevant individuals' prior well being or need, and the predicted change to their condition, and group membership.

Algorithmic fairness must represent relevant levels of well-being or need for different individuals

1. Suppose a given application is algorithmically fair.
2. If fairness is algorithmic, the elements that make a distribution fair must not be accidental. (Follows from robustness.)
3. For the elements that make a distribution fair to be non-accidental, they must be represented in the algorithm.
4. The elements that make a distribution fair are an individual's prior well-being or need, as well as [illegible]ed; as well as a representatio[illegible]
5. Therefore, the applicatio[illegible]or well being or need, and t[illegible]d group membership.

A worry about 3: What if we conceptualize fairness as *non-bias*? If that is a reasonable conception of fairness, then we don't need to represent the elements that make the distribution fair in the algorithm. Instead, we only need to ensure that there is no bias in the algorithm.

What is (algorithmic) bias?

- The training data does not represent the population to which a model is applied. (Could be due to inclusion, or annotation/ classification.) “**selection bias**”
- The model itself introduces incorrect assumptions (e.g., in terms of what it takes as input for a risk score). **“model bias”**

Annotation/ classification and bias

task instructions for data collectors and annotators. A typical assignments is illustrated by a data collection project of Active Data: the company received task instructions to collect images of diverse human faces from a Western European company, producing identification and verification systems. Eva, the founder of Active Data, offered more details:

> "They were interested in a diversity of five different ethnicities, so Caucasian, African, Middle Eastern, Latin American and Asian. Of course, very debatable whether these can be the five categories that can classify people around the world "

This type of assignment generally revolves around a client's envisioned computer vision product and underlying business idea. The technical assumptions of a classification system demand mutually exclusive categories, in this case even for a problematic concept such as race. Whether such categorisation captures the realities of data subjects or coincides with the values and believes of data workers is not negotiated. Written instructions formulated by the requester are passed along to project managers who brief workers. Workers then start collecting the images. For outsourcing companies, the rationale behind data-related decisions is "doing what the client ordered" and "offering value to the client." Conversely,

Milagros Miceli et al., "Documenting computer vision datasets: An invitation to reflexive data practices," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2021), https://dl.acm.org/doi/10.1145/3442188.3445880, p. 165

Problems with fairness as algorithmic non-bias

- Often, when the model achieves non-bias by failing to represent a given characteristic (with the idea that “justice is blind”), background inequalities are left in place.
- Sometimes, proxies for a characteristic are included in a “blind” model, either accidentally or intentionally. (Barsotti & Koçer 2024, op. cit. conclude that this is a reason to avoid such an approach.)

Suppose, then, that we try to model justice more actively. What basis should we use?

- Distributive vs. punitive justice

- Justice as fairness vs. justice as desert

- Procedural fairness vs. substantive fairness (CHLEP p. 80):

“substantive notions and rules directly refer to a particular political or normative goal or outcome a judgment or decision should effectuate”

“procedural concepts and rules describe how judgments and decisions and decisions in society should be made” (i.e., what process should be used).

Justice and mutual agreement

- In the history of Western philosophy, justice has often been identified with the basic rules that people agree to, as the basis for society.

- Contractualism (e.g., Hobbes, Locke): Justice is what people actually agree to in the “state of nature”.
- Contractarianism (e.g., Rawls, Scanlon): Justice is what reasonable people would agree to under ideal circumstances of deliberation.

Rawlsian justice as fairness

- Justice is what reasonable people would agree to as a principle of distributing freedoms/ powers and wealth, if they did not know which position in society they would go on to occupy. (This condition is known as the “Veil of Ignorance”.)
- It results in two main elements:
  - “each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others;
  - “social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone’s advantage, and (b) attached to positions and offices open to all.” (Rawls, *A Theory of Justice*, p. 60)

Three principles of egalitarianism

- The Difference Principle: After equal basic liberty and fair institutional equality are established, inequalities are permitted only when they benefit those who are least well off.
- Strict egalitarianism: Intervention (e.g., redistribution) should be used to ensure equal outcomes.
- Prioritarianism: Increases in well being for those who are less well-off *count more* than similar increases for those who are better off.

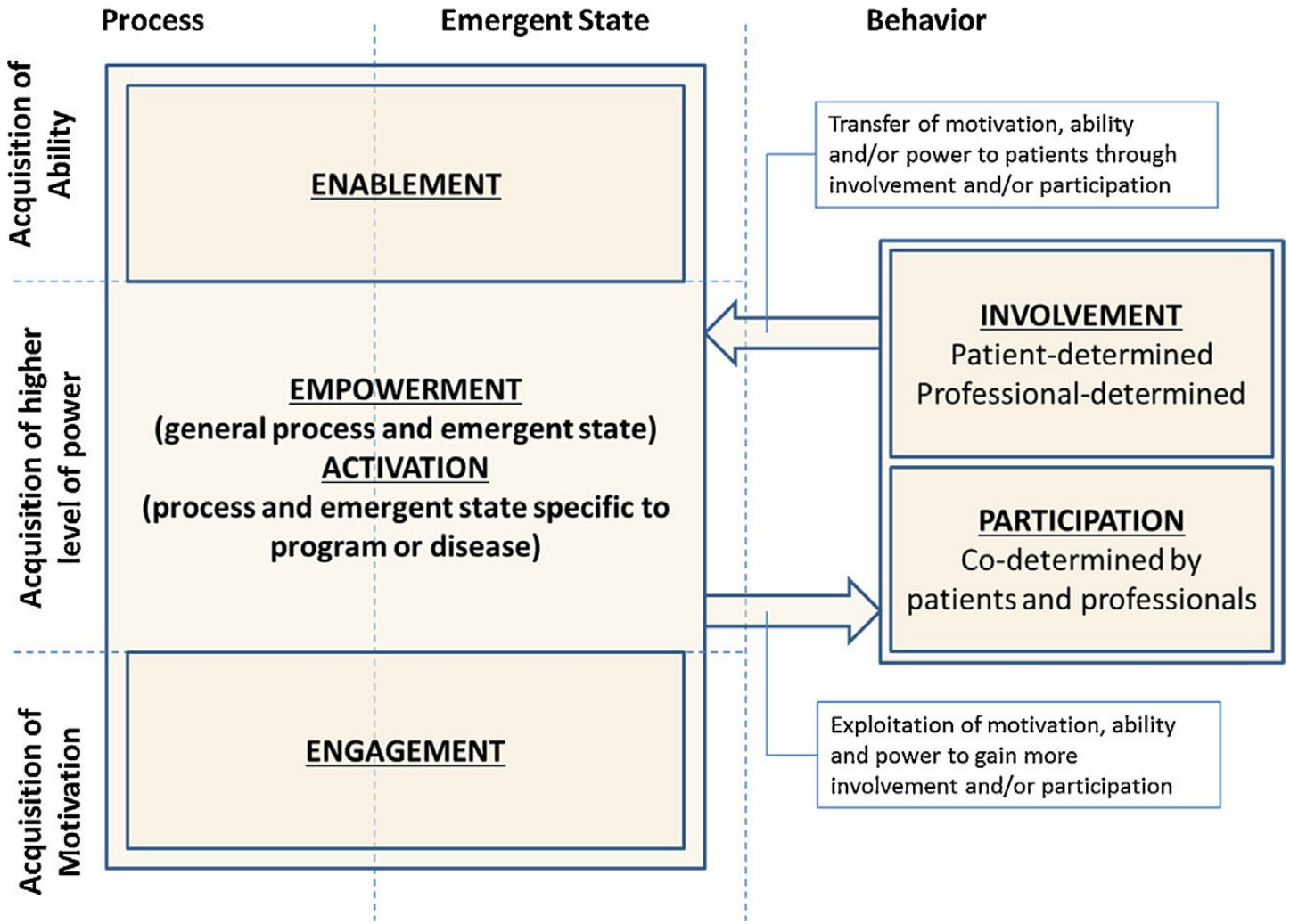
A Rawlsian rule for algorithmic fairness

- ‘A model is fair if it does not make more systematic errors for any [relevant] sub-group in the dataset compared to the others’ (Barsotti & Kocer 2024, op. cit)

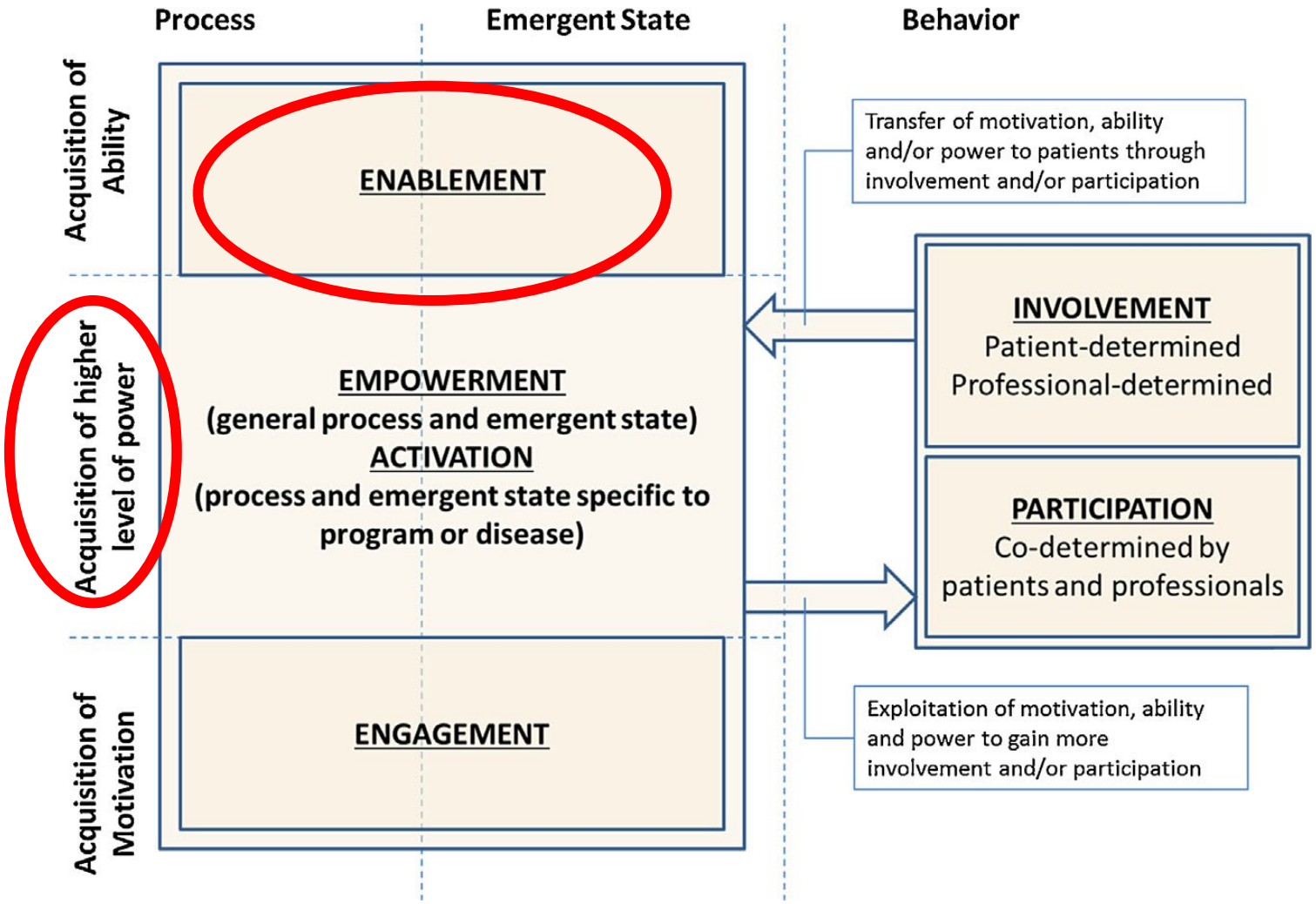
Example: digital health and justice

- Background: digital tools for health management (e.g., “persoonlijke gezondheidsomgeving” PGO) are not widely used by those with low digital literacy. It also turns out that those with low digital literacy have worse health, worse health outcomes, and are socioeconomically disadvantaged on average.
- Suppose you are designing a model that attempts to increase use of digital tools for health management in the Netherlands. You have a public health database with names, addresses, email addresses, age, and whether a person already has a PGO. The goal is to better “empower” people by increasing their awareness of health resources, by contacting them through the mail.
- What proxies for health status and benefit should be included to determine recipients, as part of a fair process?

Health empowerment in health policy literature (Fumagalli et al 2015)



Health empowerment in health policy literature (Fumagalli et al 2015)



Justice and digital health: Winters et al.'s (2020) argument

1) In the distribution of benefit, benefit should be conferred on the worse off first, even if this results in less benefit overall (Prioritarianism).
2) Therefore, in conferring health benefits, we should focus on those in poverty. (From (1))
3) In digital health, there is often a tradeoff between scaling up and general technological tools, and contextual and targeted initiatives (which benefit the poor more but require more resources).
4) In this tradeoff, we should favor contextual and targeted initiatives. (From (2) & (3))

Relational Justice

- Some approaches to justice, instead of affirming a general principle of distribution, focus instead on *power* and *history.*
- In data ethics, biobanking, and research ethics, this has led to many voicing the slogan “Nothing about us, without us” and calls for inclusivity and power-sharing among decision-makers.

HeLa (/ˈhiːlɑː/) is an immortalized cell line used in scientific research. It is the oldest human cell line and one of the most commonly used.[1][2] HeLa cells are durable and prolific, allowing for extensive applications in scientific study.[3][4] The line is derived from cervical cancer cells taken on February 8, 1951,[5] from Henrietta Lacks, a 31-year-old African American woman, after whom the line is named. Lacks died of cancer on October 4, 1951.[6]

Relational Justice and Secondary Use

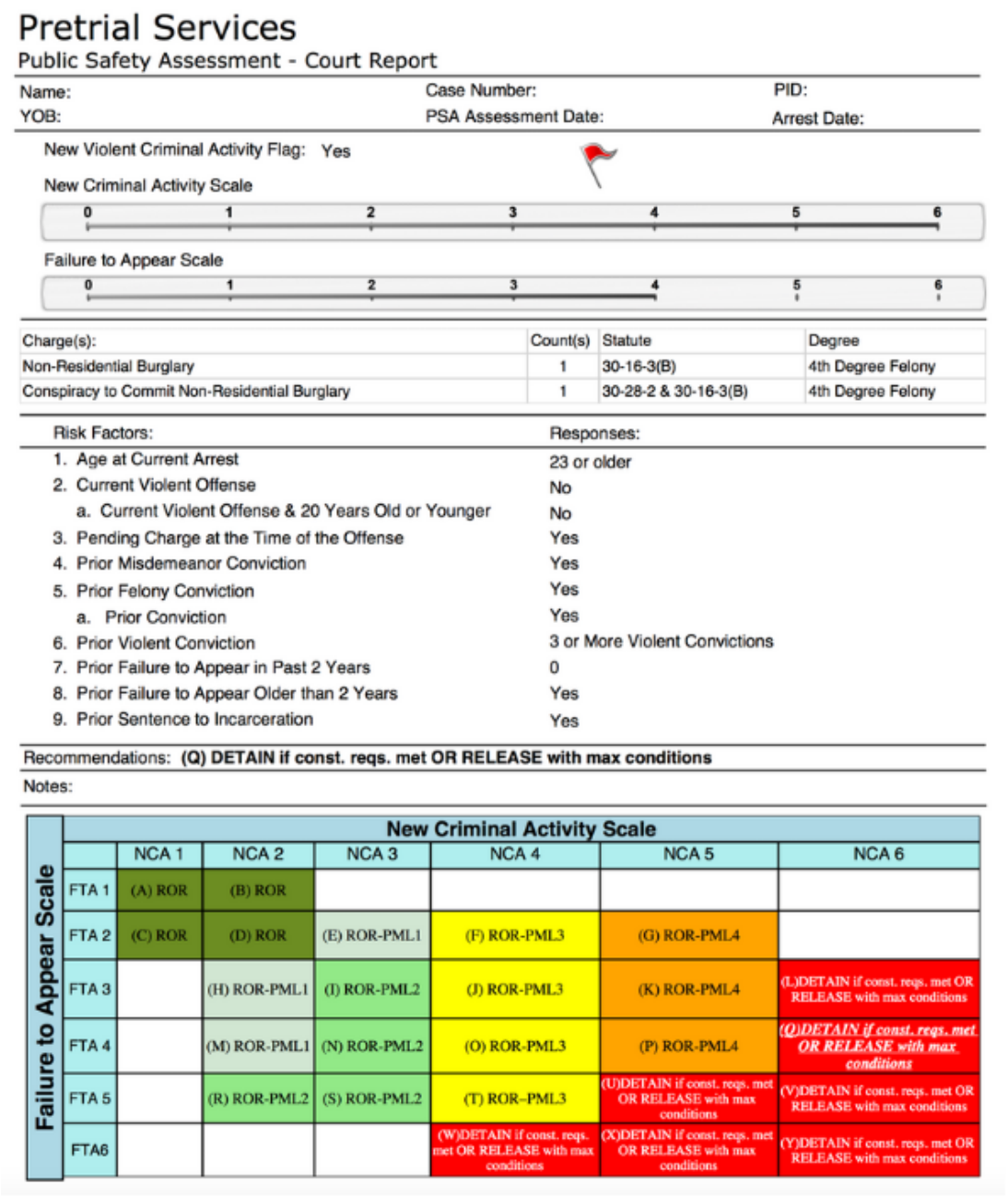
- In the early 2000s, the Havasupai tribe discovered that their DNA samples, which had been collected for studies of diabetes, had been used without consent for other genetic studies. They sued.
- A research study was conducted with IRB chairs and human genetics researchers about the case. “Interviewees perceived no direct impact from the Havasupai case on their work; if they did, it was the perceived need to safeguard themselves by obtaining broad consent or shying away from research with indigenous communities.” (Garrison 2013, 202).

Garrison, Nanibaa'A. "Genomic justice for Native Americans: impact of the Havasupai case on genetic research." *Science, Technology, & Human Values* 38, no. 2 (2013): 201-223.

Punitive Justice

Algorithms are used extensively in some parts of the world to partially automate decisions about whether to keep somebody in jail before they go to trial, or to indicate sentencing guidelines upon conviction for a crime.

Punitive justice follows a somewhat different logic from distributive justice, but it is strongly committed to the principle that “justice is blind” to arbitrary characteristics of persons such as sex, race, and income. For this reason, fairness in the sense of non-arbitrariness remains a guiding value.



Pretrial Services

Public Safety Assessment - Court Report

Name: | Case Number: | PID:

YOB: | PSA Assessment Date: | Arrest Date:

New Violent Criminal Activity Flag: Yes

New Criminal Activity Scale: 0 1 2 3 4 5 6

Failure to Appear Scale: 0 1 2 3 4 5 6

Charge(s):	Count(s)	Statute	Degree
Non-Residential Burglary	1	30-16-3(B)	4th Degree Felony
Conspiracy to Commit Non-Residential Burglary	1	30-28-2 & 30-16-3(B)	4th Degree Felony

Risk Factors:	Responses:
1. Age at Current Arrest	23 or older
2. Current Violent Offense	No
a. Current Violent Offense & 20 Years Old or Younger	No
3. Pending Charge at the Time of the Offense	Yes
4. Prior Misdemeanor Conviction	Yes
5. Prior Felony Conviction	Yes
a. Prior Conviction	Yes
6. Prior Violent Conviction	3 or More Violent Convictions
7. Prior Failure to Appear in Past 2 Years	0
8. Prior Failure to Appear Older than 2 Years	Yes
9. Prior Sentence to Incarceration	Yes

Recommendations: **(Q) DETAIN if const. reqs. met OR RELEASE with max conditions**

Notes:

Failure to Appear Scale	New Criminal Activity Scale					
	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	(A) ROR	(B) ROR				
FTA 2	(C) ROR	(D) ROR	(E) ROR-PML1	(F) ROR-PML3	(G) ROR-PML4	
FTA 3		(H) ROR-PML1	(I) ROR-PML2	(J) ROR-PML3	(K) ROR-PML4	(L)DETAIN if const. reqs. met OR RELEASE with max conditions
FTA 4		(M) ROR-PML1	(N) ROR-PML2	(O) ROR-PML3	(P) ROR-PML4	*(Q)DETAIN if const. reqs. met OR RELEASE with max conditions*
FTA 5		(R) ROR-PML2	(S) ROR-PML2	(T) ROR–PML3	(U)DETAIN if const. reqs. met OR RELEASE with max conditions	(V)DETAIN if const. reqs. met OR RELEASE with max conditions
FTA6				(W)DETAIN if const. reqs. met OR RELEASE with max conditions	(X)DETAIN if const. reqs. met OR RELEASE with max conditions	(Y)DETAIN if const. reqs. met OR RELEASE with max conditions

Figure 1

A sample Public Safety Assessment (PSA) report presented to a judge in Bernalillo County (Albuquerque) up through 2022. It shows the defendant’s Failure to Appear (FTA) and

Should we use automated scoring for punitive justice?

“So, should we use the PSA at all? Like Arnold Ventures, we do not support using the PSA to justify detention. We believe that New Mexico’s current system, where the prosecutor must present individualized evidence that a defendant is dangerous, is the right approach from an ethical and constitutional point of view. We should not deny people their liberty on statistical grounds.”

Pretrial Risk Assessment on the Ground: Algorithms, Judgments, Meaning, and Policy · Summer 2023

Engaging as AI experts

‘If you want to help the real world, dive into it. We often approach interdisciplinary work as a kind of transaction. A collaborator from another field gives us data, and we apply our techniques to it. We each stay in our lane: they are the domain experts, and we are “domain agnostic,” staying aloof from what the data is actually about.

But deep interdisciplinary work means crossing these lanes. If you care about the domain, you should learn about it. If you care about housing and credit, study the history of segregation; if you care about justice, find out how the justice system works. Approach your work with your whole brain and heart. Your life will be more intellectually satisfying, you will grow as a scientist, and your work will be more likely to be useful to the world.’

https://mit-serc.pubpub.org/pub/czviu6qc/release/2

References not linked in the text

Fumagalli, L.P., Radaelli, G., Lettieri, E., Bertele, P., Masella, C. (2015). Patient Empowerment and its neighbours: Clarifying the boundaries and their mutual relationships. *Health Policy* 119: 384-394.

Winters, N., Venkatapuram, S., Geniets, A., & Wynne-Bannister, E. (2020). Prioritarian principles of digital health in low resource settings. *Journal of Medical Ethics* 46: 259-264. doi:10.1136/medethics-2019-105468