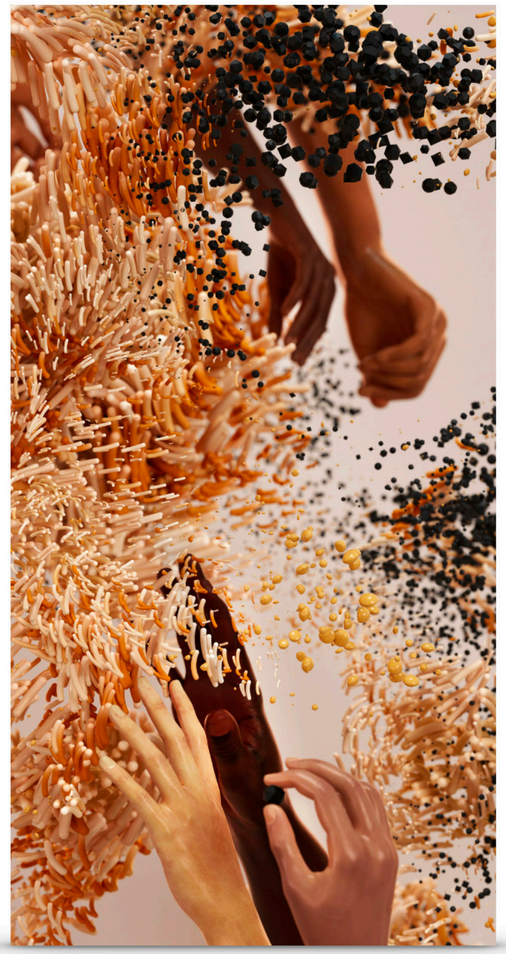





THE ROYAL SOCIETY

The Wolfson Foundation

How do we know an algorithm is treating us fairly?

Ashutosh Trivedi

University of Colorado Boulder

Royal Society Wolfson Visiting Fellow, Liverpool

CUPLV

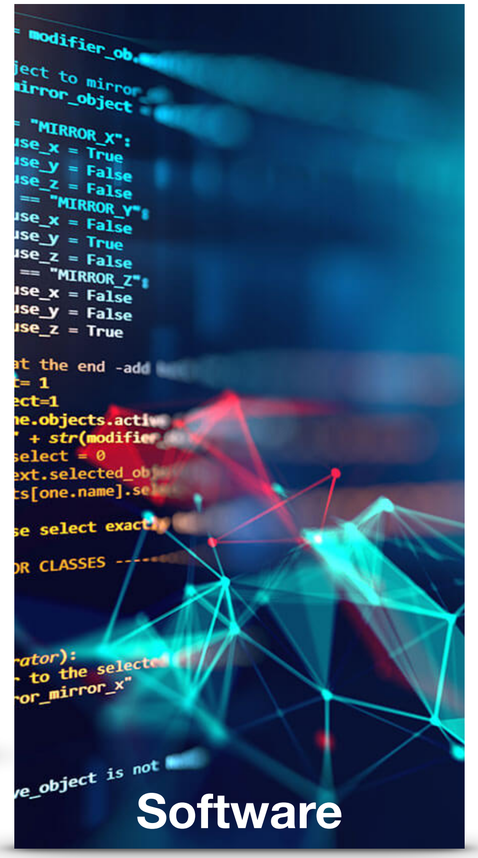
Monday, September 14 - Beyond the Bytes faculty speaker series

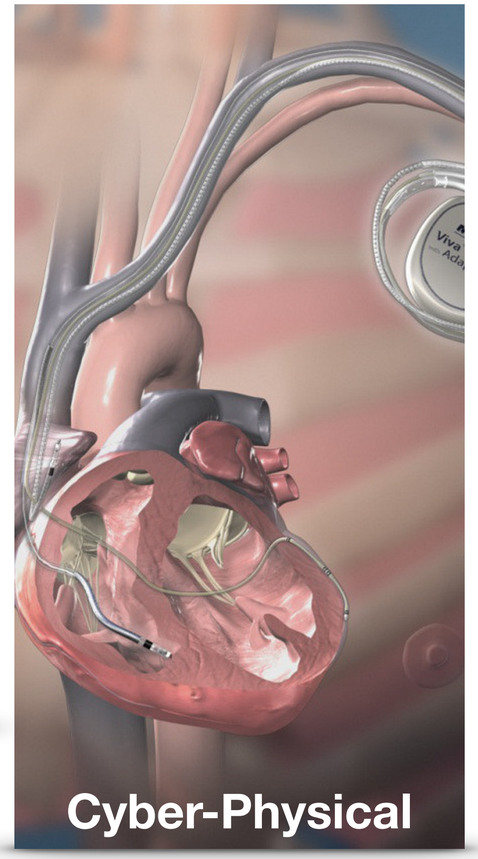
Abstract

Algorithms increasingly influence decisions about employment, credit, healthcare, taxation, and access to public services. But what does it mean for an algorithm to treat people fairly—and how can we verify that it does? Unlike traditional correctness requirements, fairness cannot usually be assessed by examining a single execution. We must compare how the algorithm behaves for different individuals or under different circumstances.

This talk introduces hyperproperties, a framework for specifying requirements that relate multiple executions of a system. Through examples from machine learning, autonomous systems, and tax-preparation software, we will see how the same perspective can capture fairness, privacy, and legal compliance. We will then explore how ideas from logic and formal verification can help determine whether a system satisfies these requirements, and why doing so is considerably more challenging than checking conventional safety properties. The talk will conclude with recent approaches that use mathematical certificates to provide guarantees about such relational behavior. No prior background in formal methods is required.







Formal Logic

Techniques

- Automata-theoretic
- Abstraction-refinement
- Inductive invariants
- Barrier certificates

Algorithms

- Symbolic model checking
- Bounded model checking
- SAT/SMT/SOS

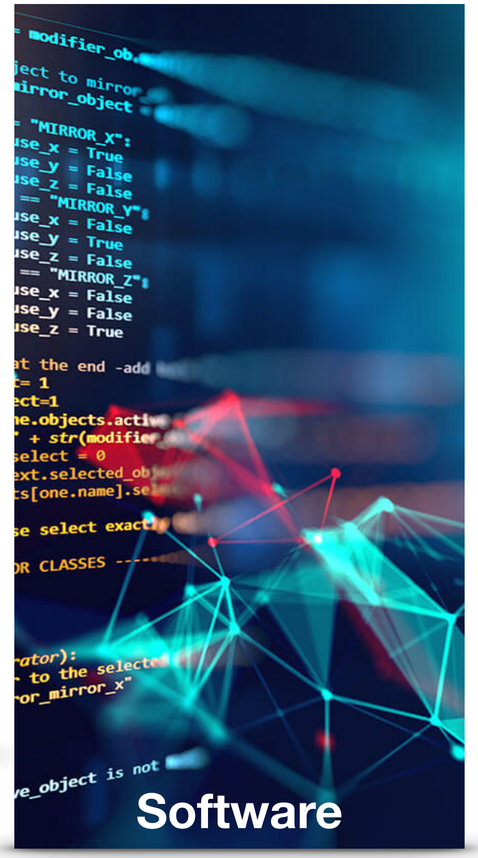
Tools

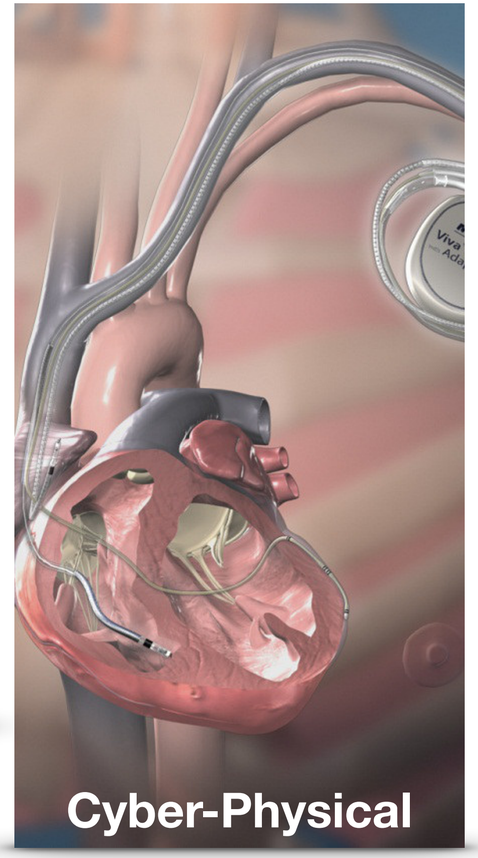
- NuSMV/PRISM/UPPAAL/..
- Phaver/DReal/Flow*
- PRISM/Storm/Owl/Mungojerrie

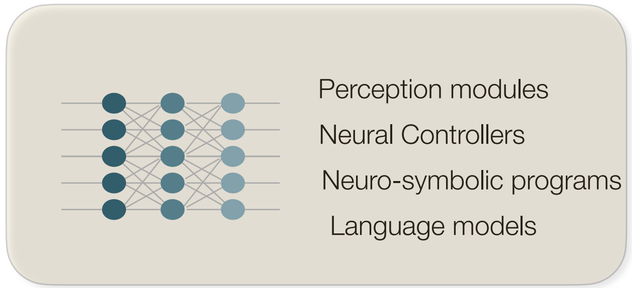
Safety-Critical Systems

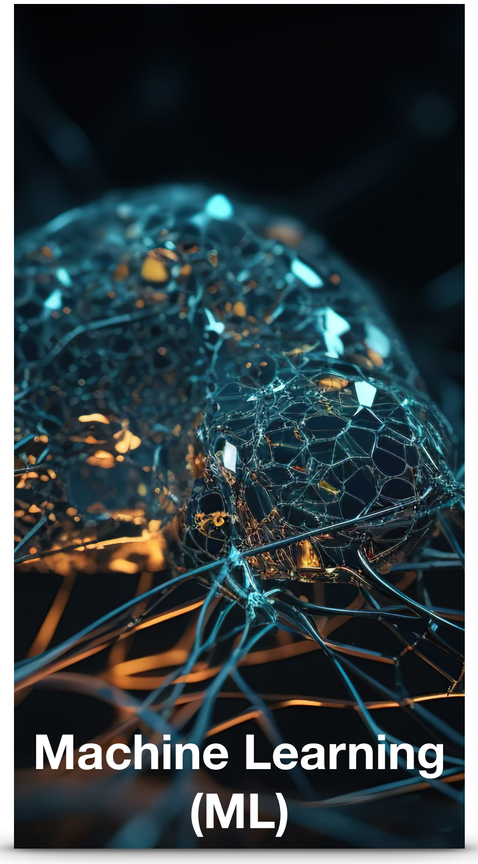
Formal Methods











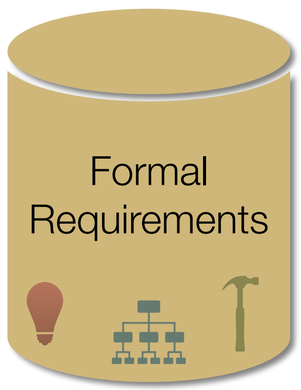
Safety-Critical Systems


Safety


Privacy


Fairness


Law



Privacy, Fairness, and Legal Requirements



Concerns beyond differential privacy: systemic leaks





Can Fam Physician. 1988 Jan; 34: 67–70.

Satisfaction and Wait Time of Patients Visiting a Family Practice Clinic

David Bestvater, Earl V. Dunn, Connie Townsend, and Wendy Nelson

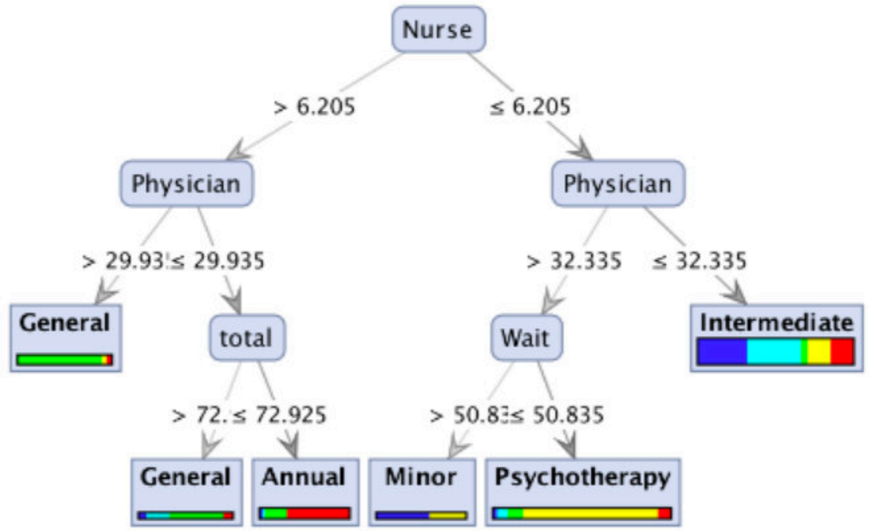
Services provided by average consultation, examination, and wait times

Service	Avg. Total Time (min)	Avg. Total Wait (min)	Avg. Time with Nurse (min)	Avg. Time with Physician (min)	No. Cases (n)	Incomplete Cases (n)
Minor assessment (std.)	50 (30)	33 (22)	1 (3)	16 (13)	67	11
Intermediate assessment (std.)	55 (24)	37 (21)	2 (3)	16 (12)	400	29
General assessment (std.)	77 (27)	31 (17)	10 (5)	36 (19)	43	1
Psychotherapy (std.)	71 (22)	35 (16)	2 (3)	34 (14)	11	0
Annual exam (after 16^{th} birthday) (std.)	51 (30)	26 (12)	7 (4)	18 (8)	5	2
Other service					13	N/A
No service code given					74	N/A

Timing can reveal information about internal behavior. The illustration motivates an observational model; it is not a general claim about every clinic.







A side-channel analysis must state what the observer measures and which inputs are secret.



Non-interference

Inputs differing only in secret values produce equivalent observations.

$\forall s_1, s_2$ and $\forall p$ we have that $\mathcal{O}(s_1, p) = \mathcal{O}(s_2, p)$

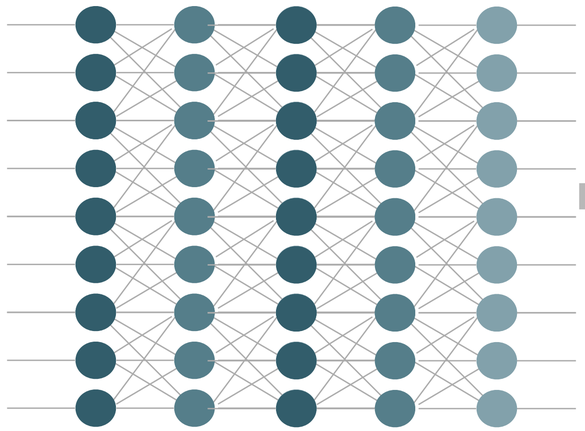
This equality states deterministic noninterference. Randomized systems need an appropriate distributional formulation.

[H] Clarkson & Schneider, Hyperproperties, JCS 2010





Historical Data



ML models



Data-driven hardware, software, or CPS

ML models may reproduce training-data biases or introduce new disparities. Testing asks whether behavior satisfies a stated fairness specification.

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023

How can we test these concerns against explicit specifications?



Fairness

WISCONSIN.GOV

State of Wisconsin
Department of Corrections

COMPAS

DOC uses the Correctional Offender Management Profiling for Alternative Sanctions tool, commonly known as COMPAS, for criminogenic
assessments and unified case planning. This actuarial risk assessment system contains offender information specifically designed to dete
and inform dynamic case plans that will guide the offender throughout his or her lifecycle in the criminal justice system.

PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Two Petty Theft Arrests

VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8

Two Drug Possession Arrests

DYLAN FUGETT	BERNARD PARKER
Prior Offense 1 attempted burglary	Prior Offense 1 resisting arrest without violence
Subsequent Offenses 3 drug possessions	Subsequent Offenses None
LOW RISK 3	HIGH RISK 10

ProPublica's main charge:
Black defendant face higher false positive rates

Historical COMPAS example (2016). Paired news examples motivate questions; they are not controlled tests with all other features fixed.

[P] Angwin et al., Machine Bias, ProPublica, May 23, 2016 · [D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023

Individual fairness as a relational specification

An attribute-invariance specification

For admissible pairs that differ only in designated protected attributes p, require $O(p_1, n) = O(p_2, n)$. Here n denotes the features held fixed.

What this tests

The model should not change its output under the particular input transformation specified by the test.

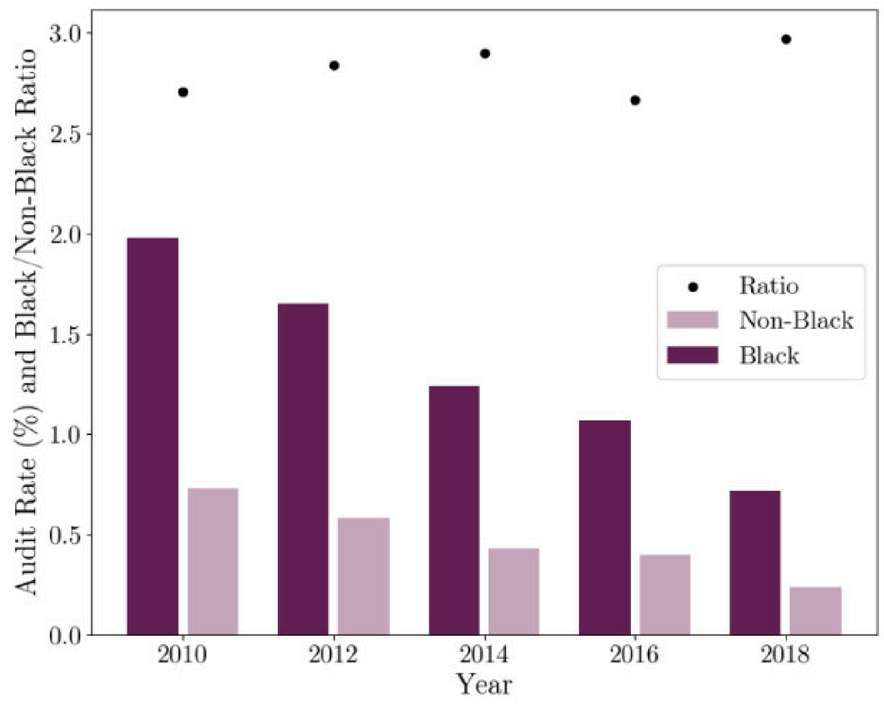
Assumptions matter

This is one formalization of individual fairness. Choosing admissible pairs and relevant qualifications requires domain knowledge. It does not settle every meaning of fairness.

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023



IRS
Department of the Treasury
Internal Revenue Service



The New York Times

Black Americans Are Much More Likely to Face Tax Audits, Study Finds

A new report documents systemic discrimination in how the I.R.S. selects taxpayers to be audited, with implications for a debate on the agency's funding.

WASHINGTON — Black taxpayers are at least three times as likely to be audited by the Internal Revenue Service as other taxpayers, even after accounting for the differences in the types o returns each group is most likely to file, a team of economists has concluded in one of the most detailed studies yet on race and the nation's tax system.

The findings do not suggest bias from individual tax enforcement agents, who do not know the race of the people they are auditing. They also do not suggest any valid reason for the I.R.S. to target Black Americans at such high rates; there is no evidence that group engages in more tax evasion than others.

Instead, the findings document discrimination in the computer algorithms the agency uses to determine who is selected for an audit, according to the study by economists from Stanford University, the University of Michigan, the University of Chicago and the Treasury Department.

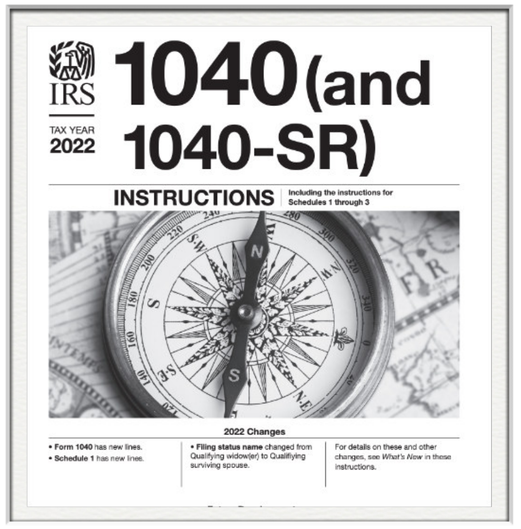
Historical motivation: the 2023 study estimates audit-rate disparities. This is distinct from a causal test on otherwise identical individuals.

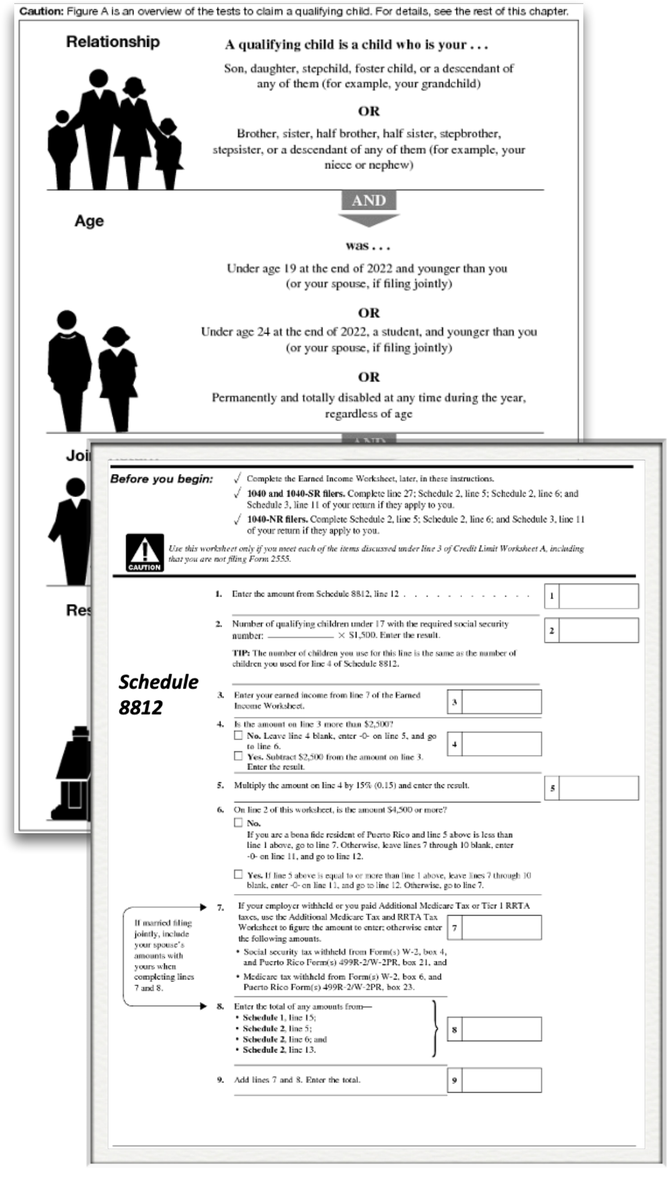
[A] Elzayn et al., Tax audit disparities, SIEPR working paper 2023

US Tax 101: Manual Filing









[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023



US Tax 101: Tax-Preparation Software





Open-source Tax Solver (OTS)

Commercial tax-preparation tools and open-source implementations provide concrete targets for software testing.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

How can software testing help detect errors in tax-preparation systems?

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

Software correctness and legal requirements

Research motivation

Tax-preparation software translates rules into computations. Defects may produce incorrect eligibility decisions or amounts.

Verification task

State the applicable rule, software version, tax year, and input assumptions before deciding that an observed output is a defect.

Scope

The examples here concern software testing. They do not establish a general rule about taxpayer liability or penalties.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023



How to detect bugs in a **legal-critical system** such as tax-preparation software?

- Absence of **oracle**
- Lack of reliable datasets
- Computationally infeasible

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023



Common Law

['kä-mən 'lȯ]

A body of unwritten laws based on legal precedents established by the courts.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

Legal specifications and metamorphic relations

Motivation from precedent

The analogy with similar legal cases motivates comparing related software inputs. A valid software relation still needs an explicit rule and its conditions.

General form

If R(x, y) holds for two admissible inputs, require S(O(x), O(y)). R specifies the input relation and S specifies the expected output relation.

Testing consequence

A violated relation supplies a candidate counterexample. Confirm the applicable rule and assumptions before classifying it as a software defect.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

Tax-Preparation Software: A Legal-Critical System





Open-source Tax Solver (OTS)

Metamorphic Testing and Debugging of Tax Preparation Software

Saeid Tizpaz-Niari
Computer Science Department
University of Texas at El Paso
saeid@utep.edu

Verya Monjezi
Computer Science Department
University of Texas at El Paso
vmonjezi@miners.utep.edu

Morgan Wagner
Psychology Department
University of Texas at El Paso
mrwagner@miners.utep.edu

Shiva Darian
Information Science Department
University of Colorado Boulder
shiva.darian@colorado.edu

Krystia Reed
Psychology Department
University of Texas at El Paso
kmreed2@utep.edu

Ashutosh Trivedi
Computer Science Department
University of Colorado Boulder
ashutosh.trivedi@colorado.edu

TENFORTY revealed three types of failures in *OpenTaxSolver*: missing some eligibility conditions (e.g., married people filing separately status is not eligible to take earned income credits); software fails to satisfy the correctness requirements when the computed tax returns get very close to zero (small non-zero values); and the updated software that allows users to explicitly opt for an option does not satisfy some correctness requirements in the corner cases.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

Hyperproperties relate multiple executions

Privacy

Noninterference: changing a secret while holding public inputs fixed preserves the observer-visible output.

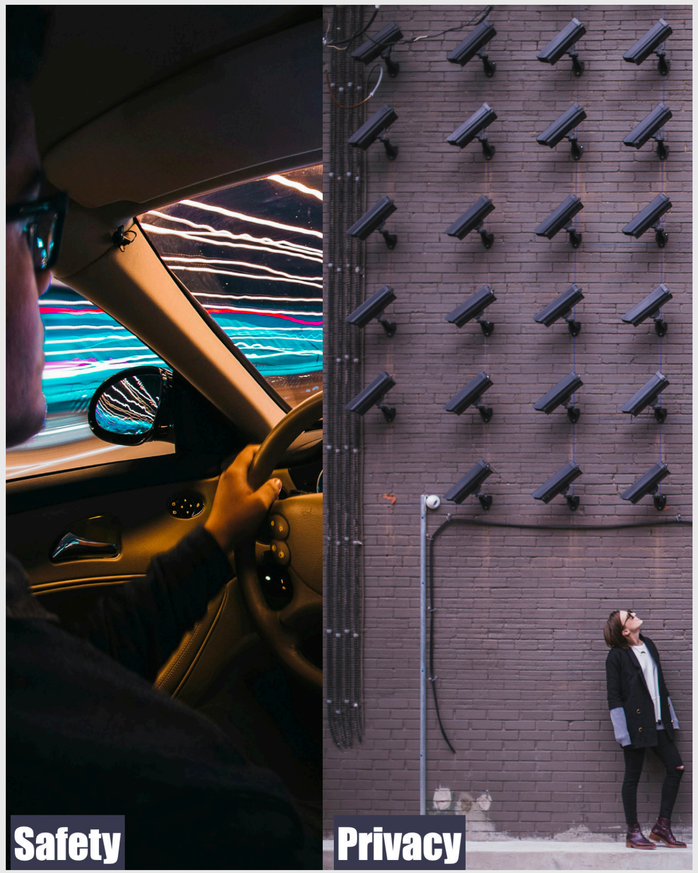
Fairness

Attribute invariance: specified changes to protected features preserve the outcome for admissible input pairs. Other fairness definitions use similarity or group statistics.

Legal requirements

Metamorphic testing: related inputs must produce outputs satisfying a specified relation. A hyperproperty is a set of sets of execution traces.

[H] Clarkson & Schneider, Hyperproperties, JCS 2010



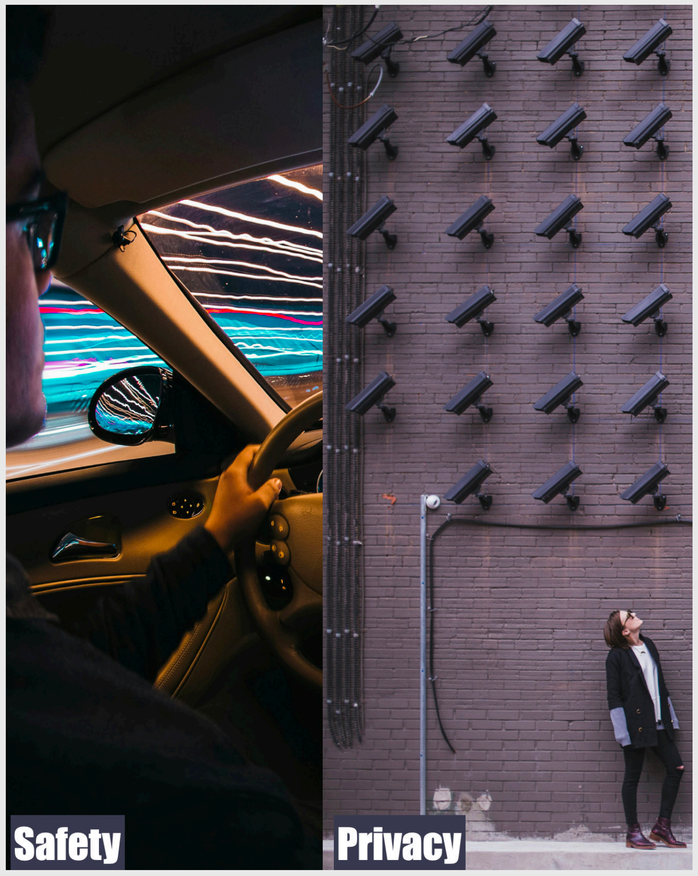
Safe and Secure CPS



Fairness Specifications



Legal Specifications



Safe and Secure CPS



Mahathi Anand
IST, Uni Stuttgart



Vishnu Murali
CU Boulder



Majid Zamani
CU Boulder

A functional certificate based approach for hyperproperty verification

Siyuan Liu, Ashutosh Trivedi, Xiang Yin, Majid Zamani:
Secure-by-construction synthesis of cyber-physical systems. Annu. Rev. Control. 53: 30-50 (2022)

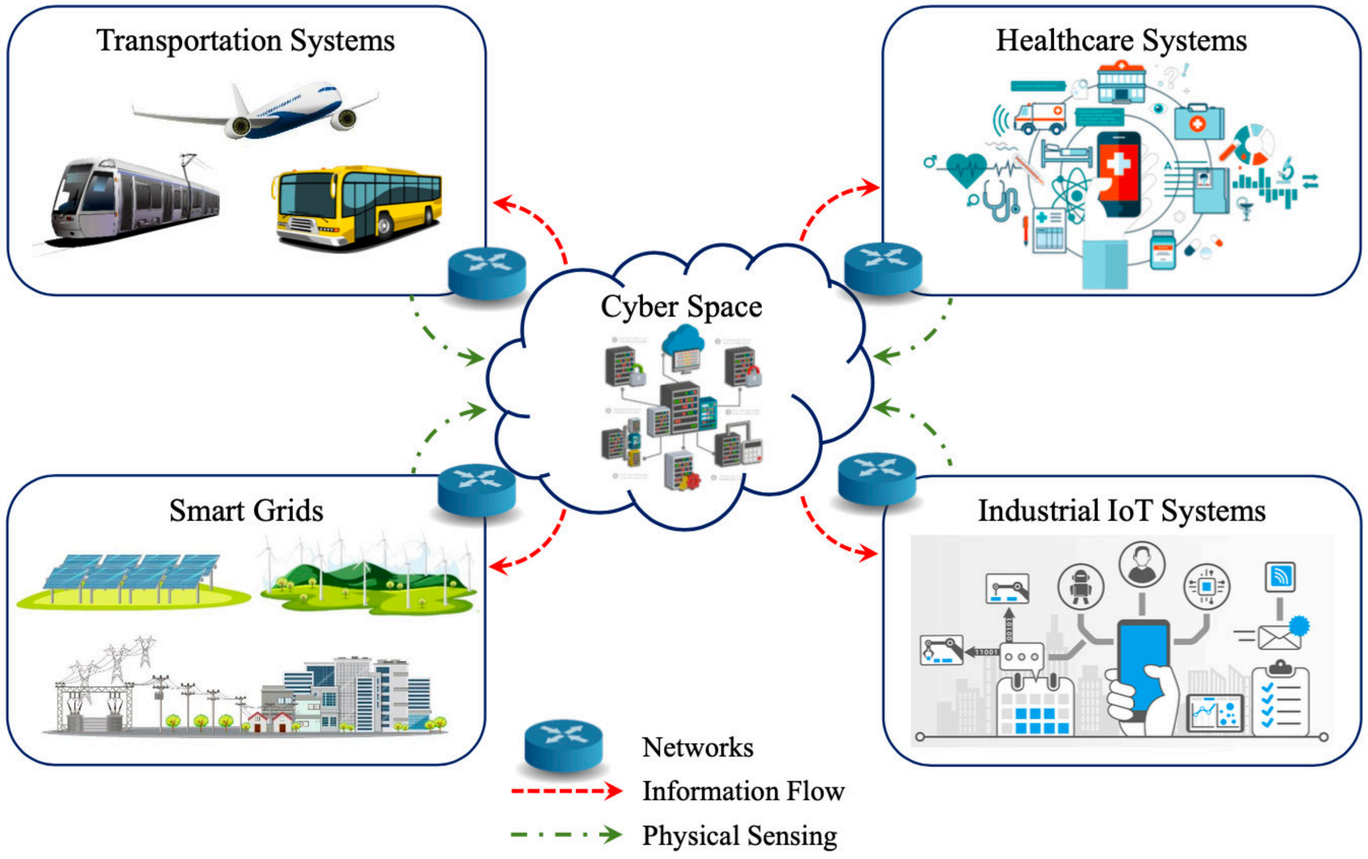
Mahathi Anand, Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Formal Verification of Control Systems against Hyperproperties via Barrier Certificates. CoRR abs/2105.05493 (2021) IEEE, TAC (2024).

Mahathi Anand, Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Formal verification of hyperproperties for control systems. CAADCPS@CPSIoTWeek 2021: 29-30

Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Closure Certificates. HSCC 2024: 10:1-10:11

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024





Cyber-Physical Systems

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024



Safety-Critical Cyber-Physical Systems

- Large-scale, interconnected, heterogeneous components
- Complex requirements
- Undecidability
- Powerful deduction and abstraction based approaches

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

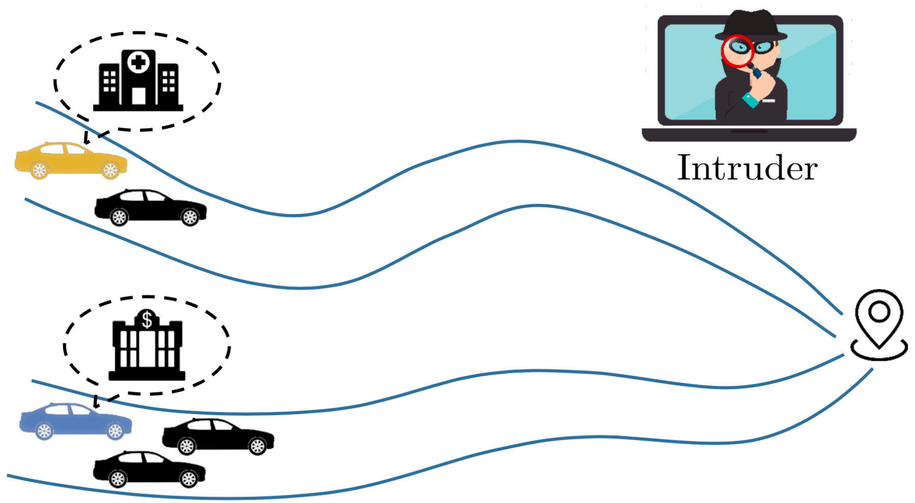


Security-Critical Cyber-Physical Systems

- Novel attack surfaces
- Information leakage in physical variable
- Undecidability
- Rich history of security analysis

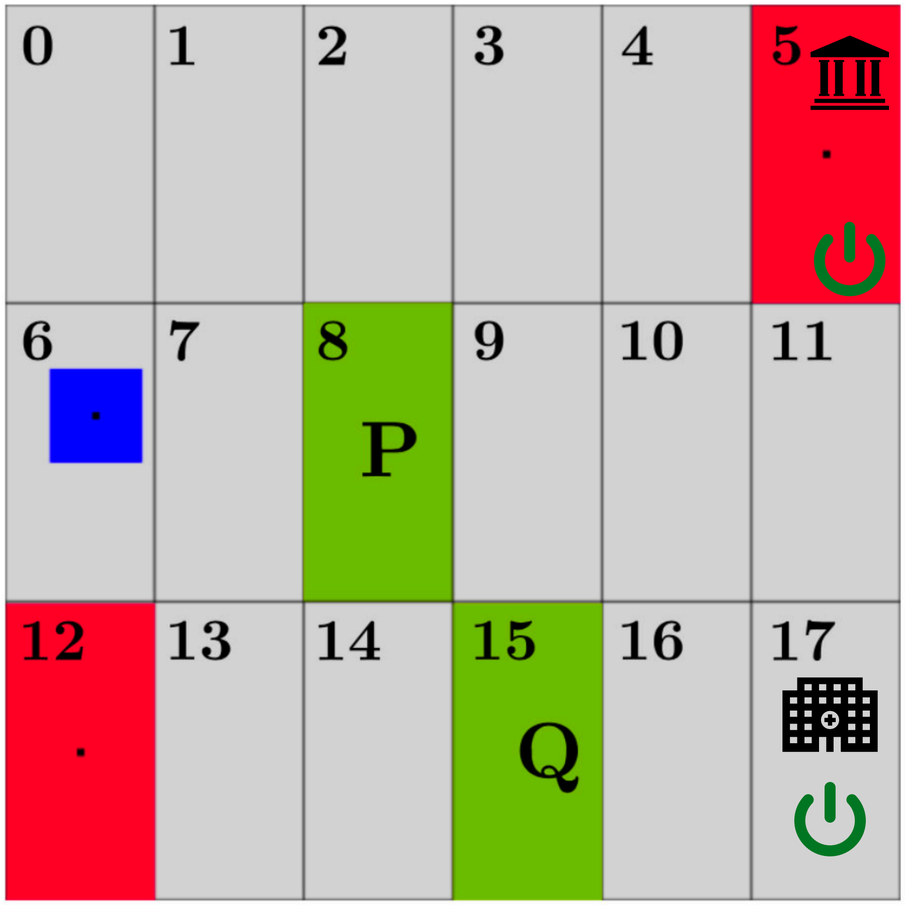
  (Entropy, opacity, hyperproperties)

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

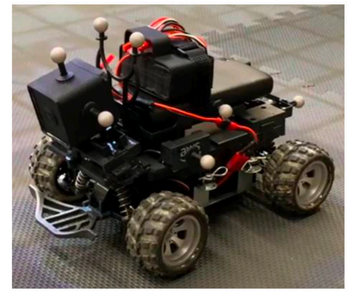


Attack Model

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024



$$\Sigma : \begin{cases} \dot{x} = v\cos(\theta) \\ \dot{y} = v\sin(\theta) \\ \dot{\theta} = \frac{v}{L}\tan(u_1) \\ \dot{v} = av + bu_2 \end{cases}$$



Mission-Objective: The vehicle visits Parking (P) and Quay (Q) locations infinitely often.

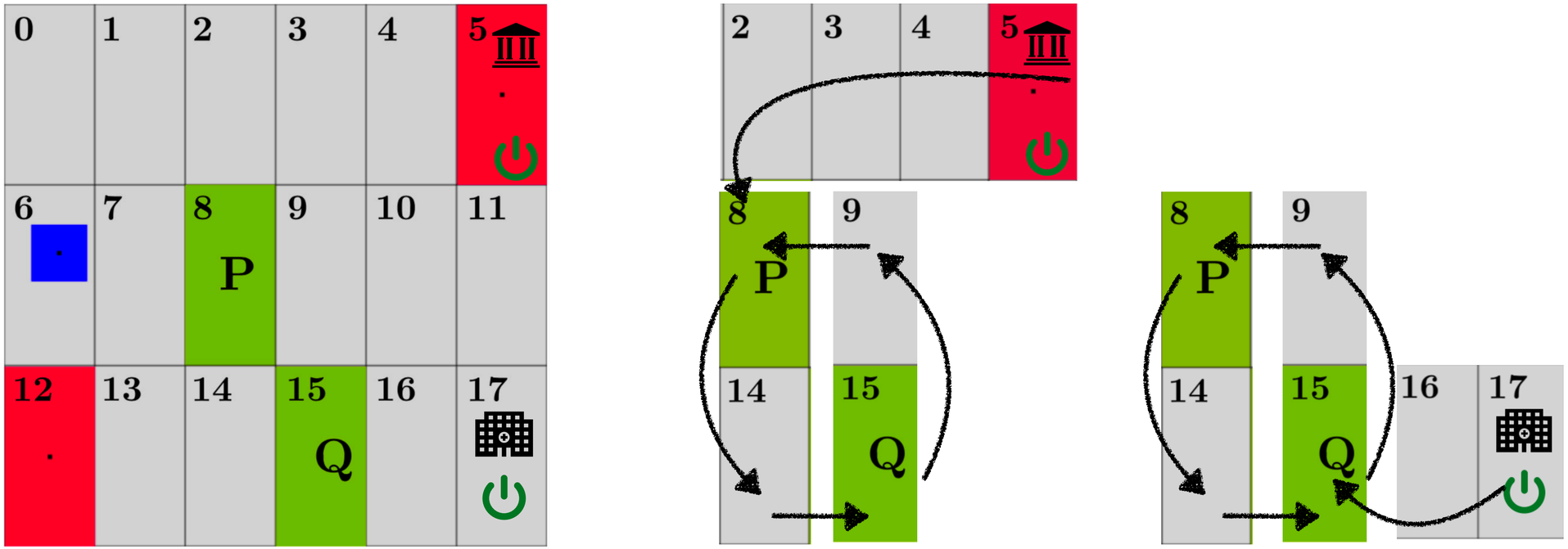
$$\forall\pi(\square\Diamond P_\pi) \quad \wedge \quad (\square\Diamond Q_\pi)$$

Security-Objective (Opacity): The intruder must not infer whether the vehicle started from a secret location.

$$\forall\pi\exists\pi'\left((\mathtt{Init}_\pi \wedge \mathtt{Init}_{\pi'} \wedge \mathtt{Secret}_\pi) \implies \neg\mathtt{Secret}_{\pi'} \wedge G(\mathcal{O}_\pi \iff \mathcal{O}_{\pi'})\right)$$

Trace variables range over valid initialized executions. The existential witness must start non-secret whenever the universal trace starts secret.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024



Mission-Objective: The vehicle visits Parking (P) and Quay (Q) locations infinitely often.

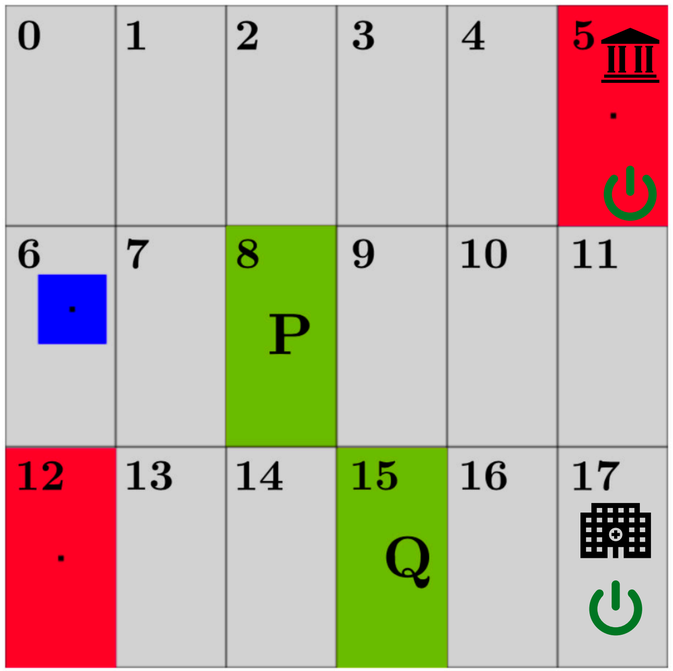
$$\forall \pi (\square \Diamond P_\pi) \quad \wedge \quad (\square \Diamond Q_\pi)$$

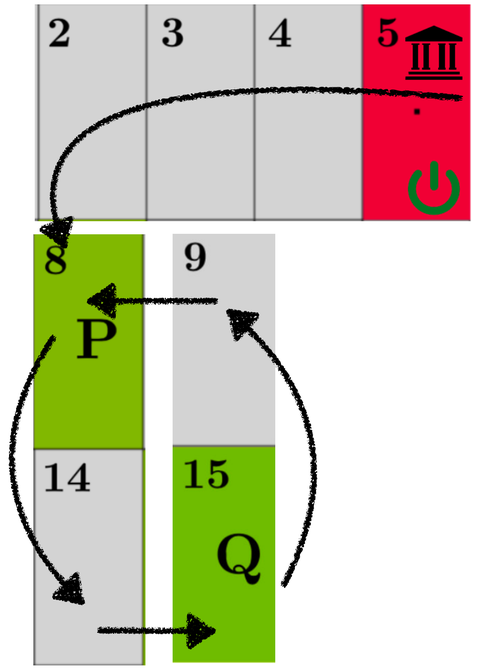
Security-Objective (Opacity): The intruder must not infer whether the vehicle started from a secret location.

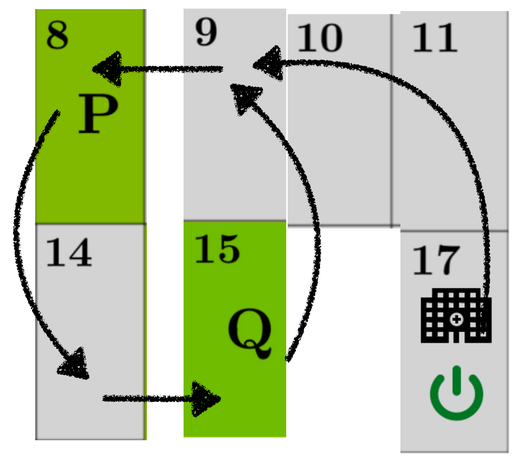
$$\forall \pi \exists \pi' \left((\mathtt{Init}_\pi \wedge \mathtt{Init}_{\pi'} \wedge \mathtt{Secret}_\pi) \implies \neg \mathtt{Secret}_{\pi'} \wedge G(\mathcal{O}_\pi \iff \mathcal{O}_{\pi'}) \right)$$

Opacity requires a non-secret witness with the same observations; a successful mission alone does not establish opacity.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024







Mission-Objective: The vehicle visits Parking (P) and Quay (Q) locations infinitely often.

$$\forall \pi (\square \Diamond P_\pi) \quad \wedge \quad (\square \Diamond Q_\pi)$$

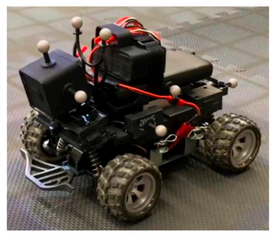
Security-Objective (Opacity): The intruder must not infer whether the vehicle started from a secret location.

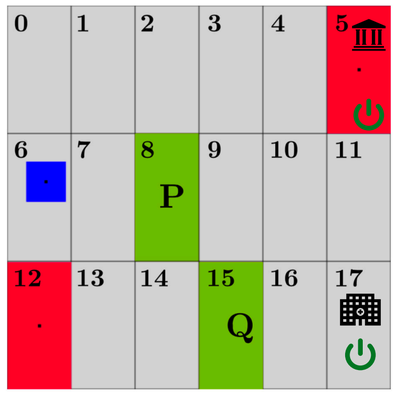
$$\forall \pi \exists \pi' \left((\mathtt{Init}_\pi \wedge \mathtt{Init}_{\pi'} \wedge \mathtt{Secret}_\pi) \implies \neg \mathtt{Secret}_{\pi'} \wedge G(\mathcal{O}_\pi \iff \mathcal{O}_{\pi'}) \right)$$

The mission and the information-flow requirement are separate proof obligations.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

$$\Sigma : \begin{cases} \dot{x} = v\cos(\theta) \\ \dot{y} = v\sin(\theta) \\ \dot{\theta} = \frac{v}{L}\tan(u_1) \\ \dot{v} = av + bu_2 \end{cases}$$





$$\forall\pi(\square\Diamond P_\pi) \quad \wedge \quad (\square\Diamond Q_\pi)$$

$$\forall\pi\exists\pi'\left((\mathtt{Init}_\pi \wedge \mathtt{Init}_{\pi'} \wedge \mathtt{Secret}_\pi) \implies \neg\mathtt{Secret}_{\pi'} \wedge G(\mathcal{O}_\pi \iff \mathcal{O}_{\pi'})\right)$$

Problem: HyperLTL Verification

Given a discrete-time dynamical system S and a hyperLTL φ specification, decide whether the system satisfies the specification, i.e. $S \models \varphi$.

- Undecidable
- Classic abstractions don’t preserve opacity
- Opacity preserving abstractions
- Deductive verification (functional certificates)

Mahathi Anand, Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Formal Verification of Control Systems against Hyperproperties via Barrier Certificates. CoRR abs/2105.05493 (2021) IEEE, TAC (2024).

Rajeev Alur, Pavol Cerný, Steve Zdancewic:
Preserving Secrecy Under Refinement. ICALP (2) 2006: 107-118

Xiang Yin, Majid Zamani, Siyuan Liu:
On Approximate Opacity of Cyber-Physical Systems. IEEE Trans. Autom. Control. 66(4): 1630-1645 (2021)

The general infinite-state problem is undecidable. Ordinary abstractions need not preserve opacity.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

Verifying CPS using functional certificates

- Barrier certificates (inductive state-invariants)
- Lyapunov certificates (ranking functions)
- Closure certificates (inductive transition invariants)

Stephen Prajna, Ali Jadbabaie:
Safety Verification of Hybrid Systems Using Barrier Certificates. HSCC 2004: 477-492

Antonis Papachristodoulou, Stephen Prajna:
On the construction of Lyapunov functions using the sum of squares decomposition. CDC 2002: 3482-3487

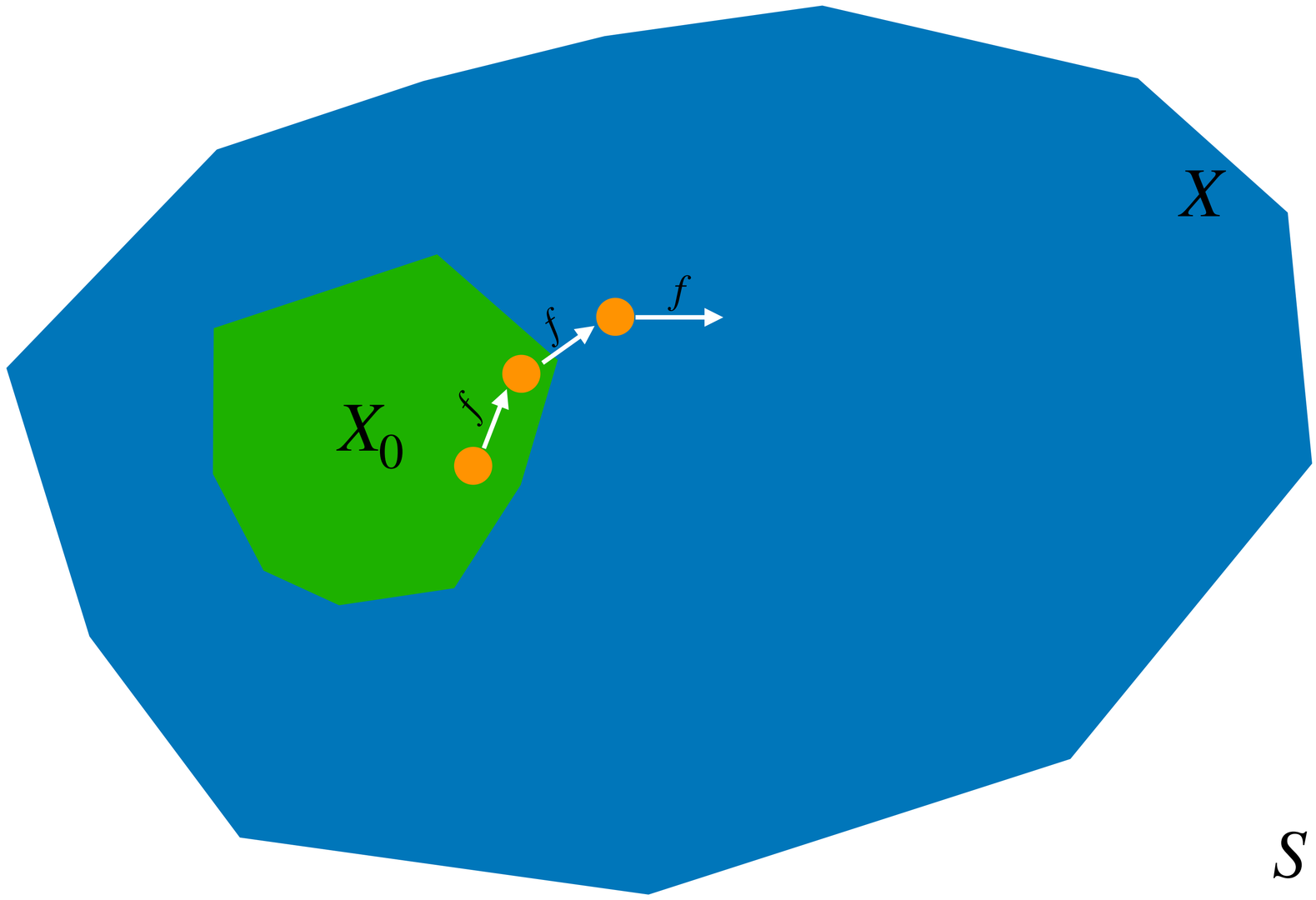
Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Closure Certificates. HSCC 2024: 10:1-10:11

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024

Verifying Safety using Barrier Certificates

Stephen Prajna, Ali Jadbabaie:
Safety Verification of Hybrid Systems Using Barrier Certificates. HSCC 2004: 477-492

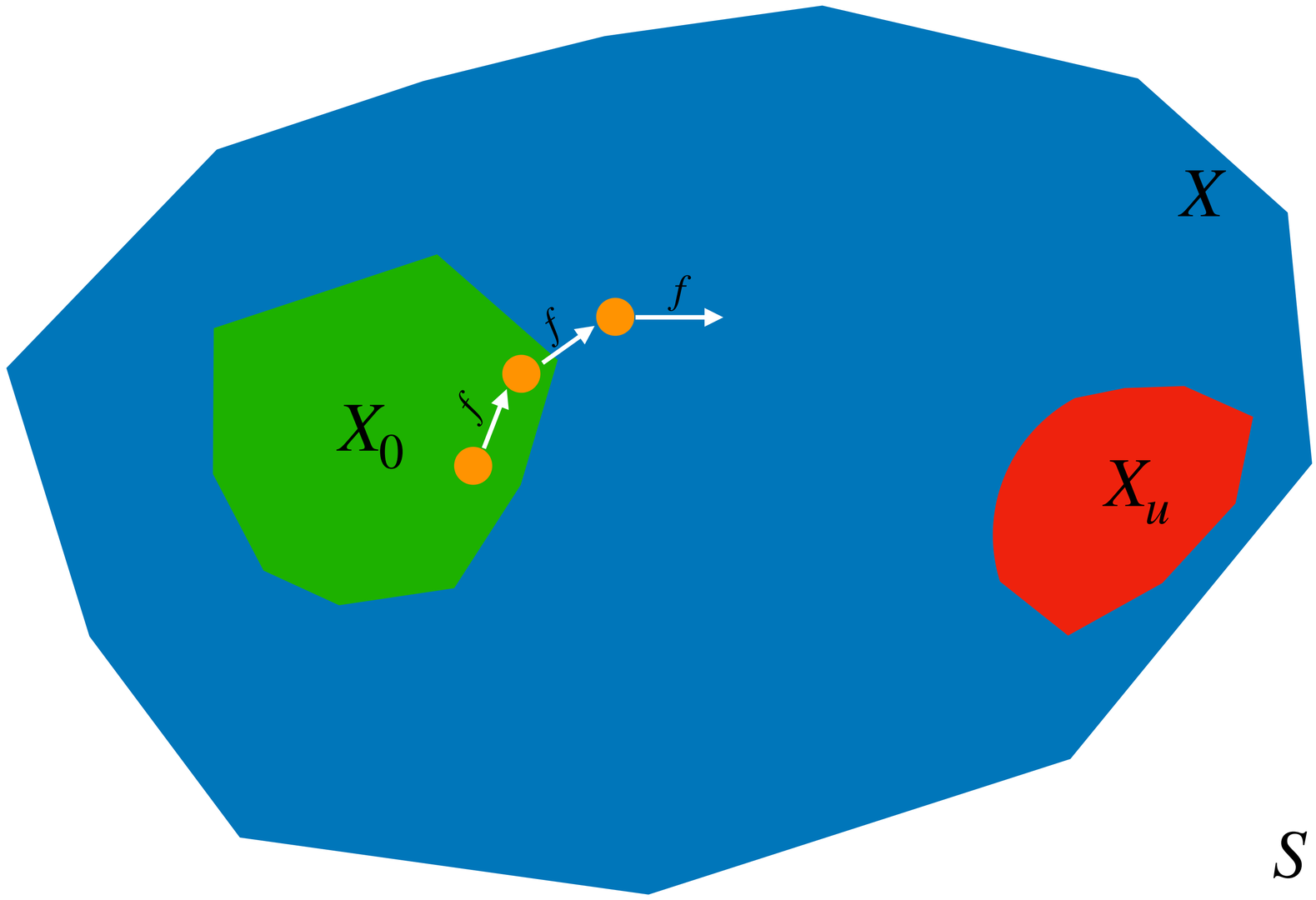
[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024



$$S = (X, X_0, f)$$

Discrete-Time Dynamical Systems

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024



Unsafe set: X_u

$$S = (X, X_0, f)$$

Safety Verification

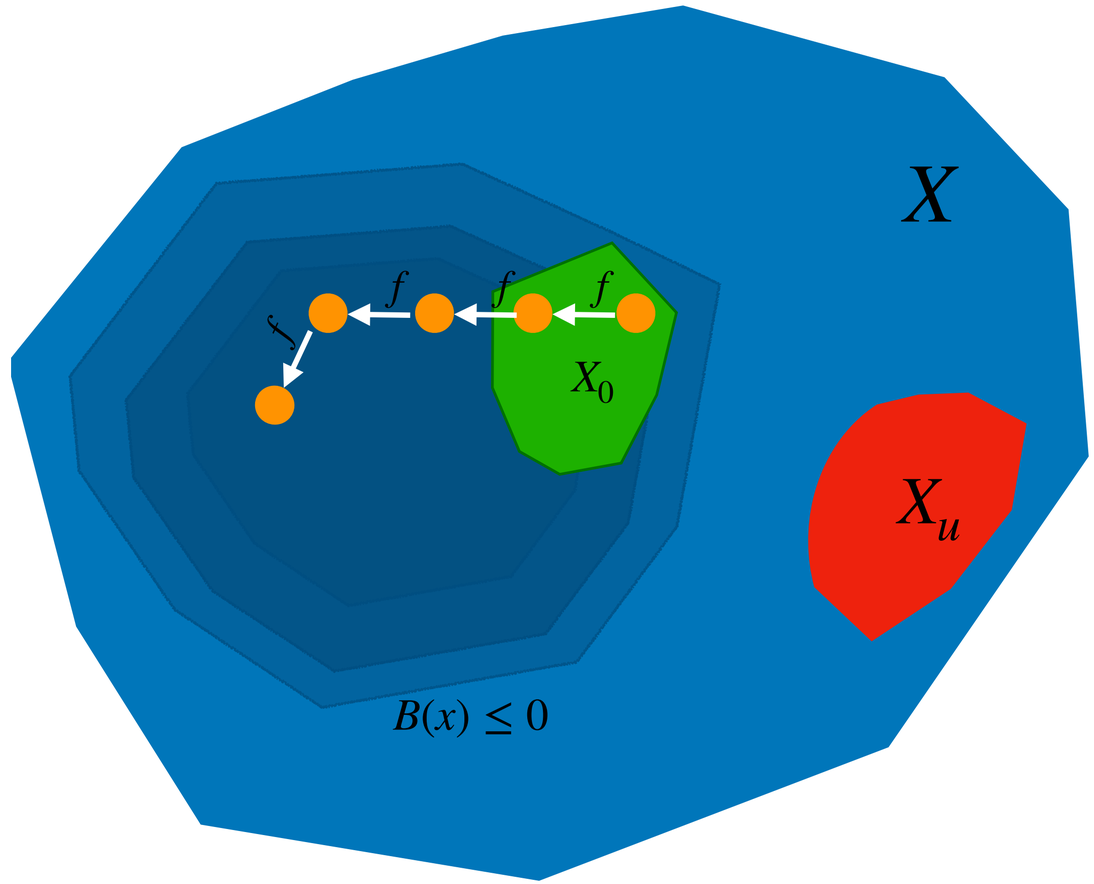
Discrete-Time Dynamical Systems

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024

Barrier Certificates

Sufficient conditions for safety
B: X → $\mathbb{R}$

B(x) ≤ 0 for x ∈ X_0
B(x) > 0 for x ∈ X_u
B(f(x)) ≤ B(x) for every x ∈ X \ X_u



The sublevel set {x : B(x) ≤ 0} is an inductive invariant.

Theorem: Barrier Certificates and Safety.

1. The existence of a barrier certificate implies safety.

2. A number of methods exist for discovering *barrier certificates* including **sum-of-square** (SOS) relaxation, **optimization**, and counterexample-guided inductive synthesis, and machine learning.

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024

$$\forall\pi\exists\pi'\left((\texttt{Init}_{\pi}\wedge\texttt{Init}_{\pi'}\wedge\ \texttt{Secret}_{\pi})\implies\neg\texttt{Secret}_{\pi'}\wedge G(\mathcal{O}_{\pi}\iff\mathcal{O}_{\pi'})\right)$$

Verifying Hypersafety using Barrier Certificates

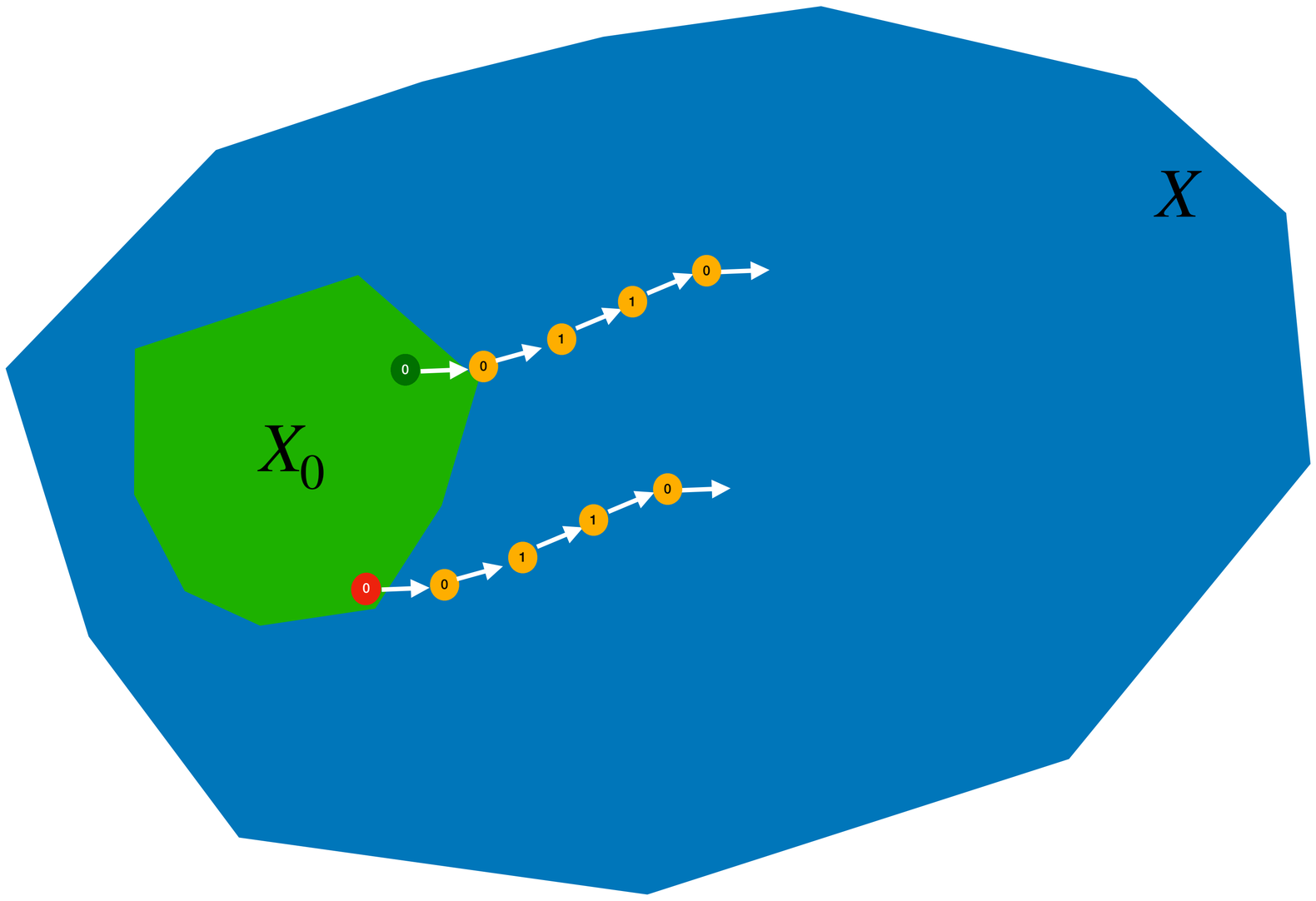
Mahathi Anand, Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Formal Verification of Control Systems against Hyperproperties via Barrier Certificates. CoRR abs/2105.05493 (2021) IEEE, TAC (2024).

Mahathi Anand, Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Formal verification of hyperproperties for control systems. CAADCPS@CPSIoTWeek 2021: 29-30

Opacity is the hyperproperty illustrated here; do not infer that every ∀∃ HyperLTL formula is hypersafety.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

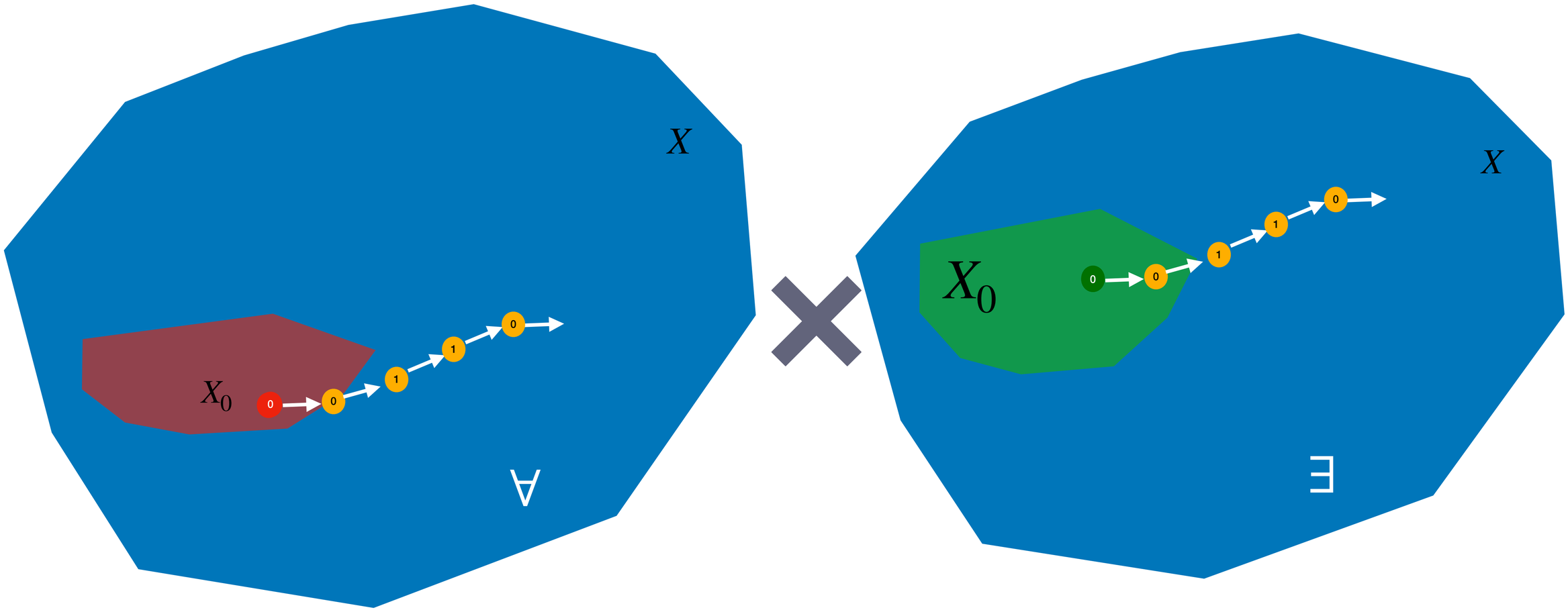
$$\forall \pi \exists \pi' \left((\mathtt{Init}_{\pi} \wedge \mathtt{Init}_{\pi'} \wedge \ \mathtt{Secret}_{\pi}) \implies \neg \mathtt{Secret}_{\pi'} \wedge G(\mathcal{O}_{\pi} \iff \mathcal{O}_{\pi'}) \right)$$



Read the displayed opacity formula with both traces ranging over initialized executions.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

$$\forall \pi \exists \pi' \left((\mathtt{Init}_\pi \wedge \mathtt{Init}_{\pi'} \wedge \mathtt{Secret}_\pi) \implies \neg \mathtt{Secret}_{\pi'} \wedge G(\mathcal{O}_\pi \iff \mathcal{O}_{\pi'}) \right)$$

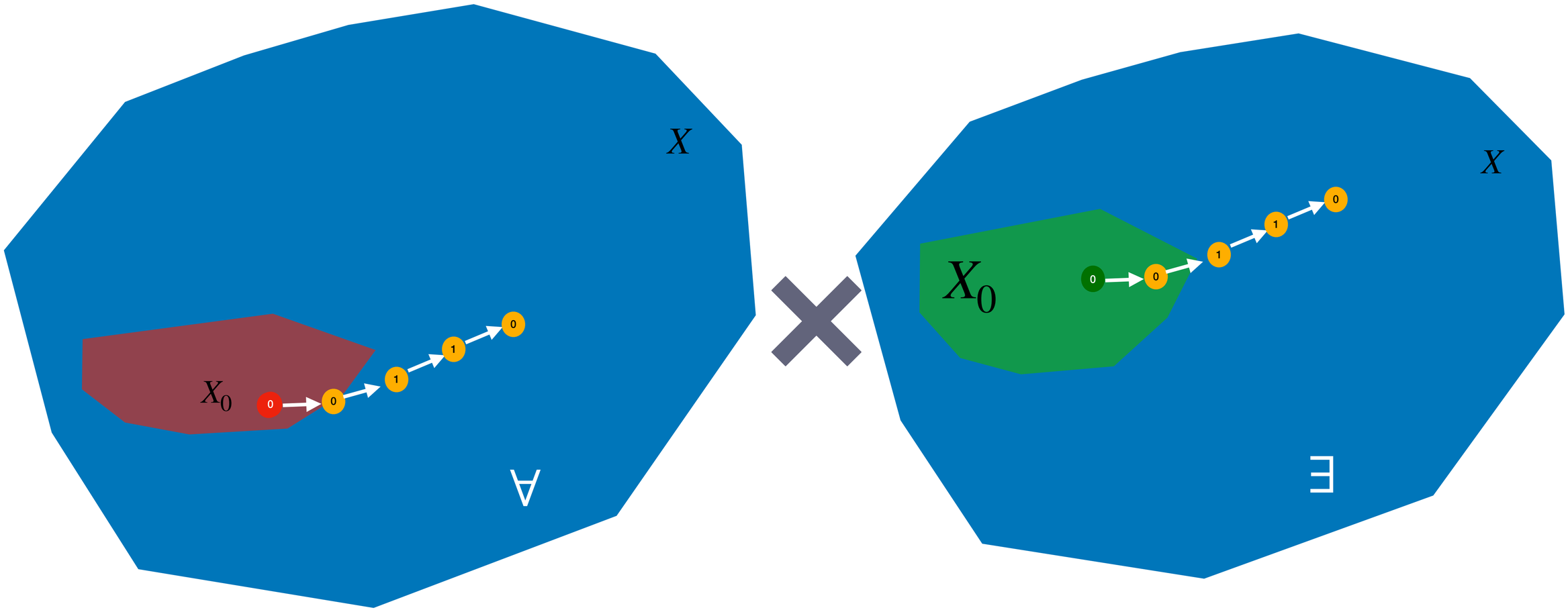


HyperLTL semantics

Standard ∀∃ trace semantics allows the witness to depend on the entire universal trace.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

$$\forall \pi \exists \pi' \left((\mathtt{Init}_{\pi} \wedge \mathtt{Init}_{\pi'} \wedge \mathtt{Secret}_{\pi}) \implies \neg \mathtt{Secret}_{\pi'} \wedge G(\mathcal{O}_{\pi} \iff \mathcal{O}_{\pi'}) \right)$$



Turn-based semantics

Winning strategy for ∃-player under turn-based semantics implies one under standard semantics.

The turn-based strategy supplies a sufficient witness construction; the converse need not hold.

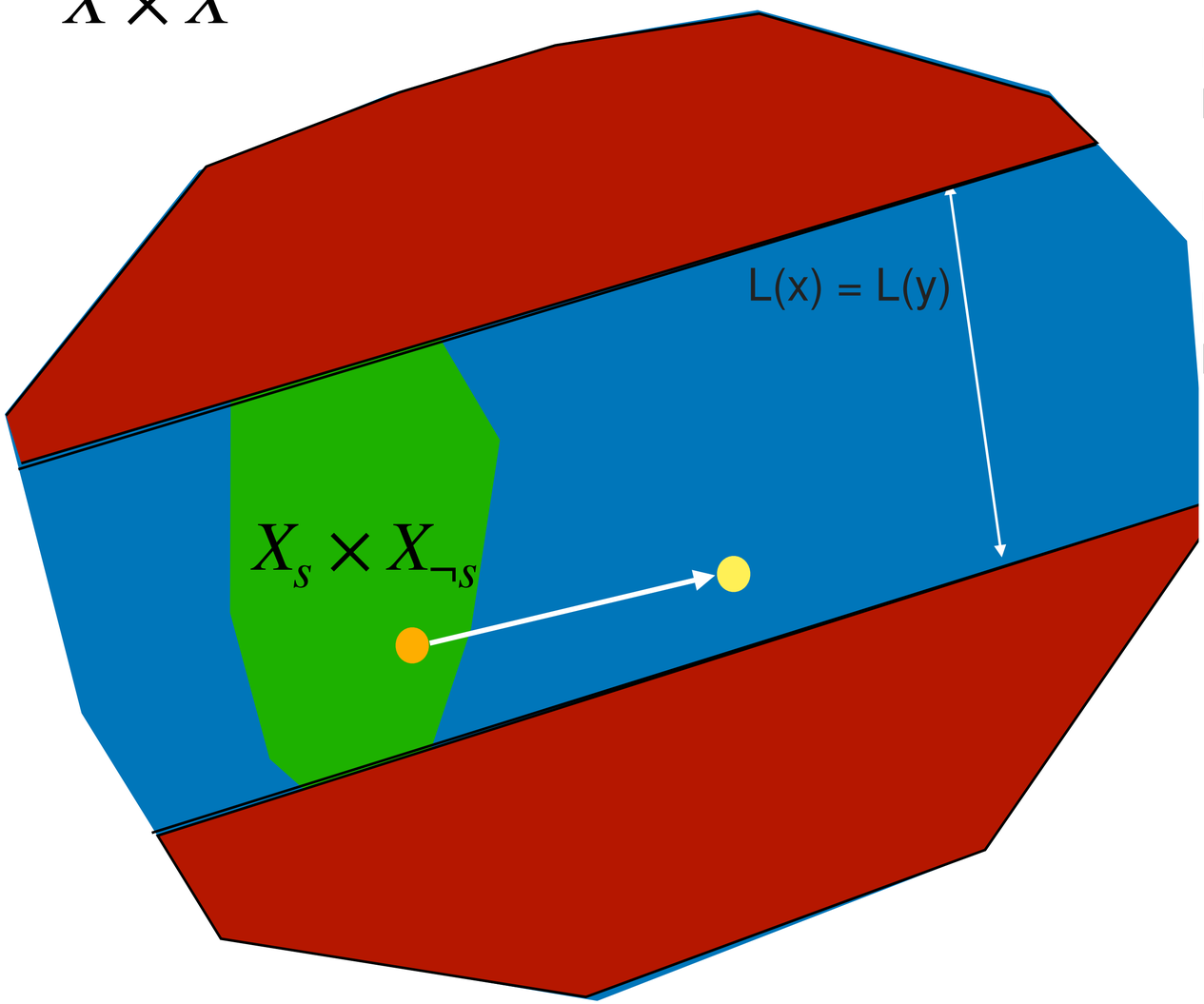
[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

Augmented barrier certificates (ABCs)

Idea. On the self-composition, for every state-pair with matching observations (**safe** state-pairs), ∃-player can match ∀-player's move to keep the system in such safe states.

More precisely, the certificate preserves a chosen invariant subset of state pairs with matching observations.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024



Certificate for initial-state opacity
B: X × X → ℝ

For every secret initial x, there is a non-secret initial y with B(x,y) ≤ 0.

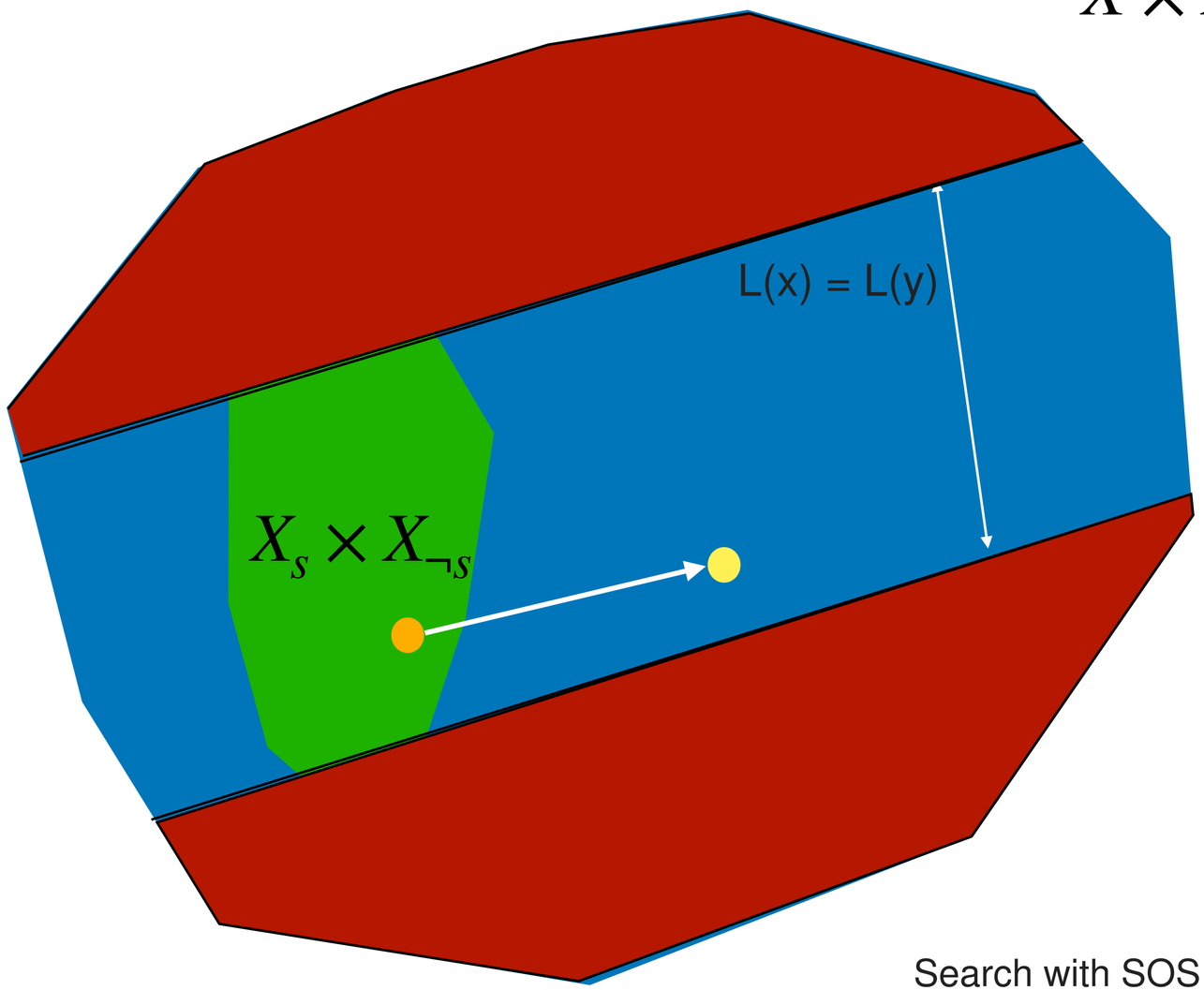
If L(x) ≠ L(y), then B(x,y) > 0.

If B(x,y) ≤ 0, then for every $x' \in F(x)$, there is $y' \in F(y)$ with B(x′,y′) ≤ 0.

Here F(x) is the set of successors and L is the observation map. Each secret initial state needs at least one non-secret matching witness.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

Augmented Barrier Certificates



Certificate for initial-state opacity
B: X × X → ℝ

For every secret initial x, there is a non-secret initial y with B(x,y) ≤ 0.

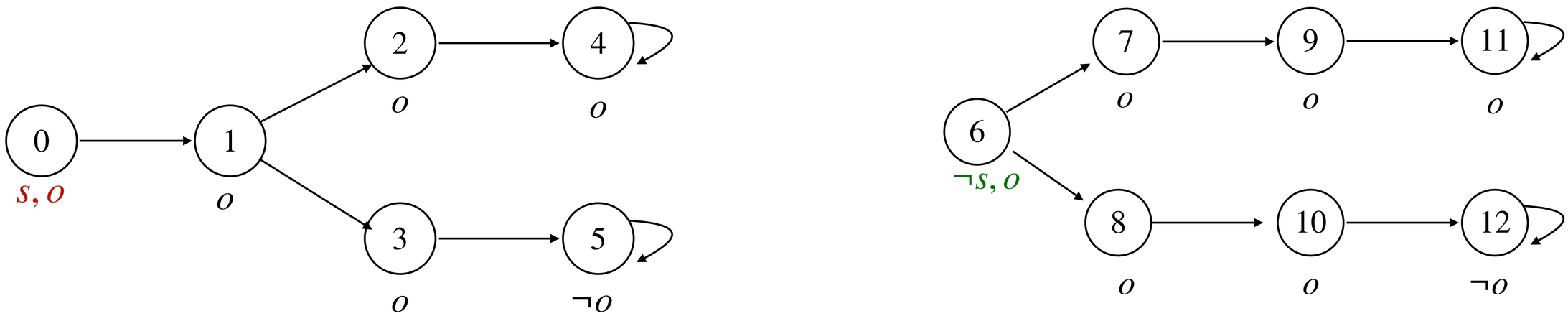
If L(x) ≠ L(y), then B(x,y) > 0.

If B(x,y) ≤ 0, then for every $x' \in F(x)$, there is $y' \in F(y)$ with B(x′,y′) ≤ B(x,y).

Search with SOS relaxations or CEGIS; failure to find a certificate is inconclusive.

The nonincrease condition is a sufficient strengthening of invariant preservation. Synthesis failure does not refute opacity.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024



Augmented barrier certificates approach is not complete.

∃-player may require an arbitrary lookahead to choose his actions.
Can we capture this long term connectivity using function certificates?

Idea. Transition-closure and Closure Certificates

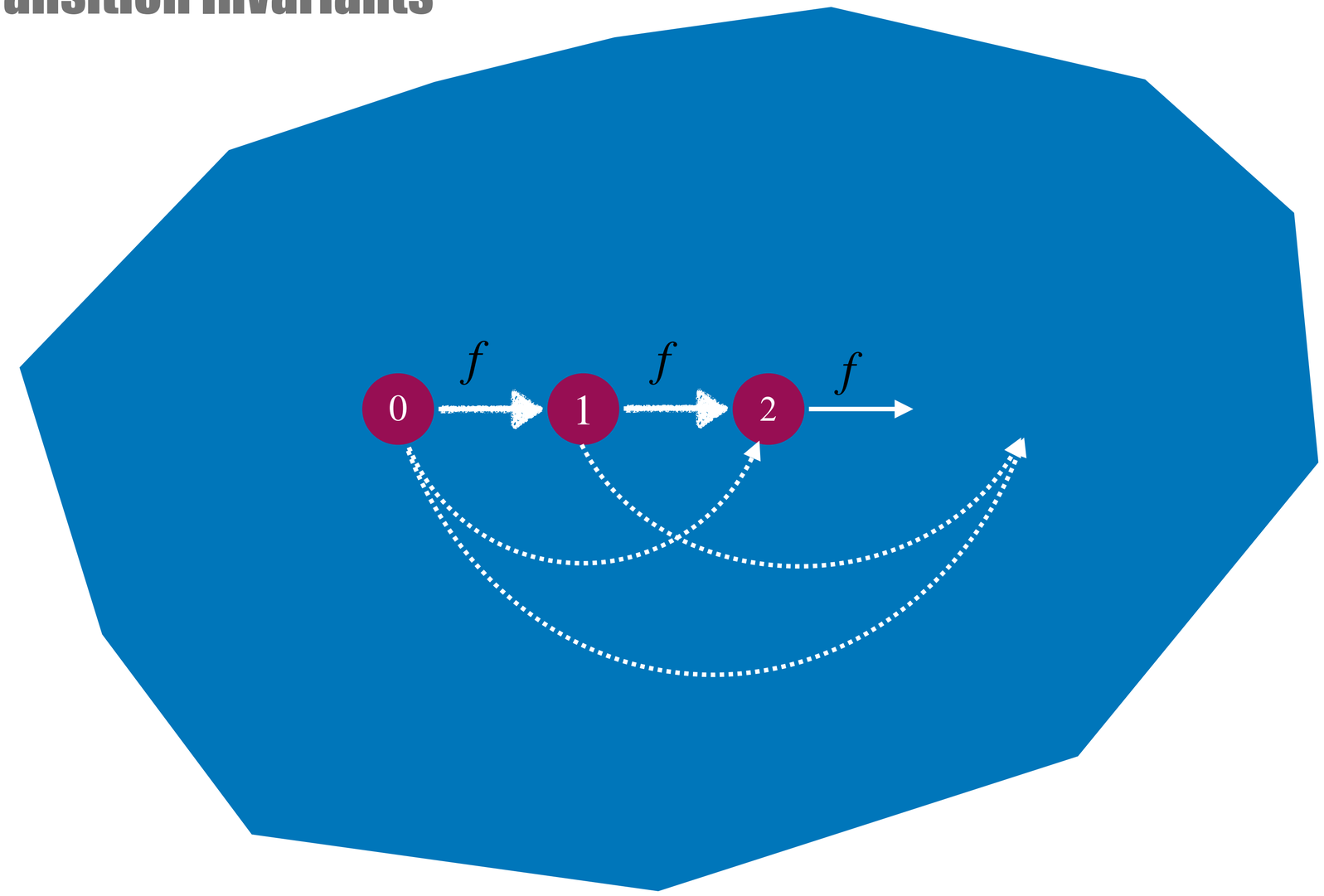
This example explains a limitation of step-by-step witness strategies. See the HyperCertificates preprint for related developments.

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

Inductive Transition Invariants and Closure Certificates

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024

Inductive Transition Invariants



For any $x \in X, (x, f(x)) \in Tinv$

If $(f(x), y) \in Tinv$, then $(x, y) \in Tinv$

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024

Closure Certificates: Finite Visit

Choose a fixed δ > 0. These conditions certify only finitely many visits to X_{VF}.

$$T : X \times X \to \mathbb{R}$$

For all $x, y \in X$

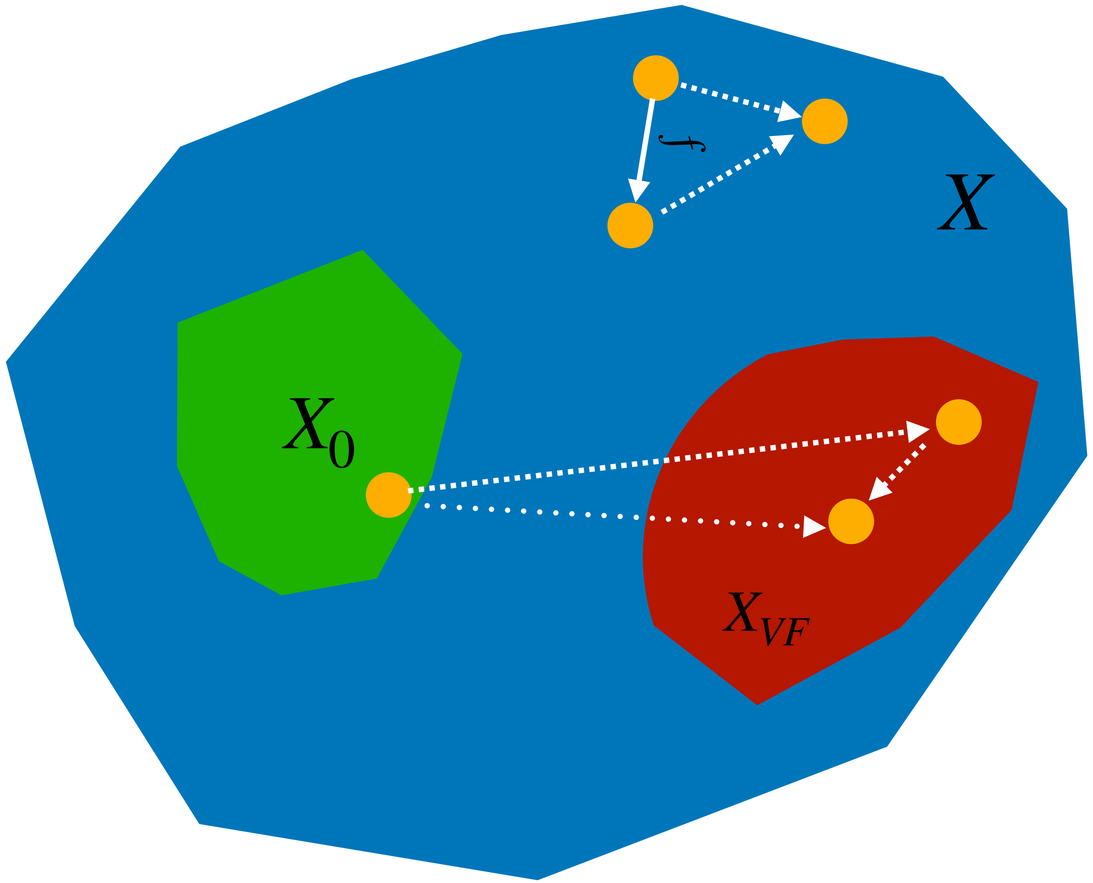
$$T(x, f(x)) \geq 0$$

$$T(f(x), y) \geq 0 \text{ , implies } T(x, y) \geq 0 \text{ ,}$$

For all $x_0 \in X_0, y, y' \in X_{VF}$

$$T(x_0, y) \geq 0 \text{ , and } T(y, y') \geq 0, \text{ implies}$$

$$T(x_0, y') \leq T(x_0, y) - \delta$$



δ is a fixed positive constant. The decrease condition, together with nonnegative reachability values, rules out infinitely many visits.

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024

Transition invariants for opacity

Why lookahead matters

A witness trace may depend on future choices of the universally quantified trace. A step-by-step matching strategy can therefore be too restrictive.

Research question

Can transition-closure information express the needed lookahead and support a certificate proof? Completeness depends on the semantics and allowed certificate class.

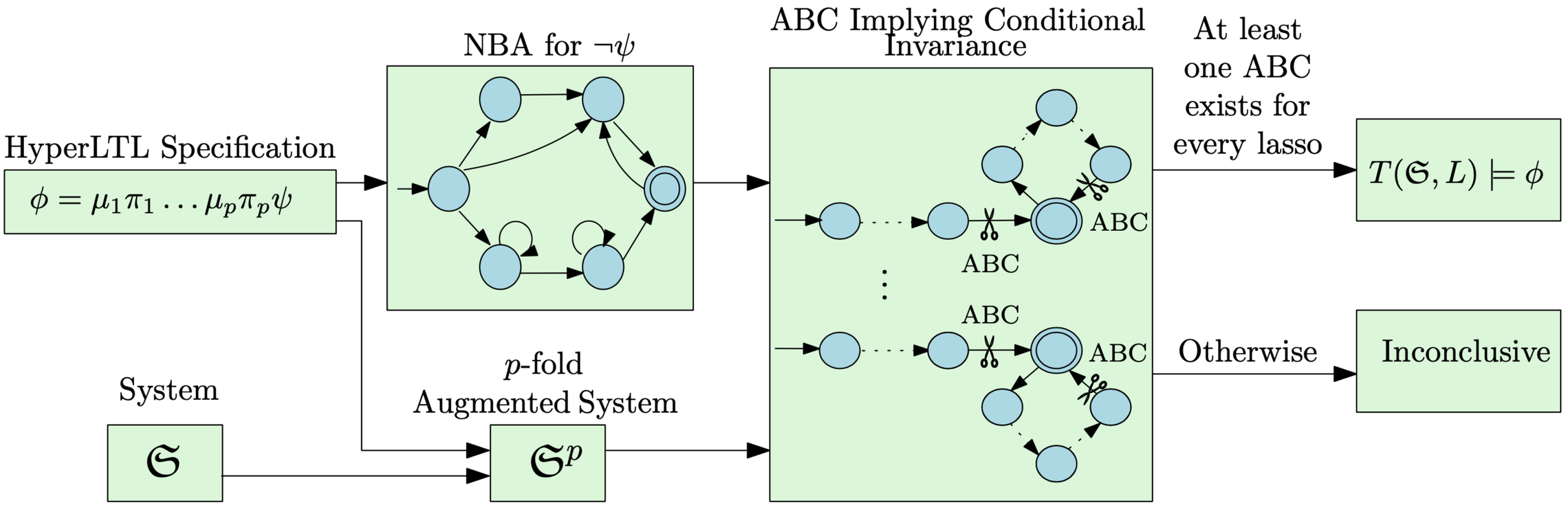
Reading connection

Closure Certificates develops transition invariants for temporal properties. HyperCertificates (2026 preprint) develops this direction for HyperLTL. The two results have different scopes.

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024 · [X] Murali, Falah, Trivedi & Zamani, HyperCertificates, 2026 preprint

Verifying Hyperproperties using Barrier Certificates

[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024



Mahathi Anand, Vishnu Murali, Ashutosh Trivedi, Majid Zamani:
Formal Verification of Control Systems against Hyperproperties via Barrier Certificates. CoRR abs/2105.05493 (2021) IEEE, TAC (2024).

Tichakorn Wongpiromsarn, Ufuk Topcu, Andrew G. Lamperski:
Automata Theory Meets Barrier Certificates: Temporal Logic Verification of Nonlinear Systems. IEEE Trans. Autom. Control. 61(11): 3344-3355 (2016)

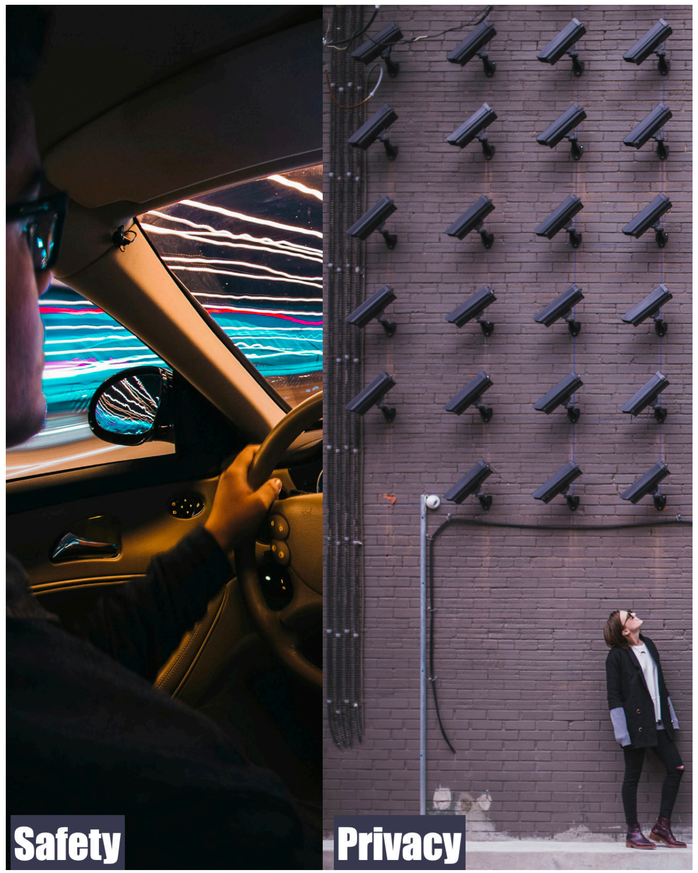
[B] Anand, Murali, Trivedi & Zamani, IEEE TAC 2024

Future Directions

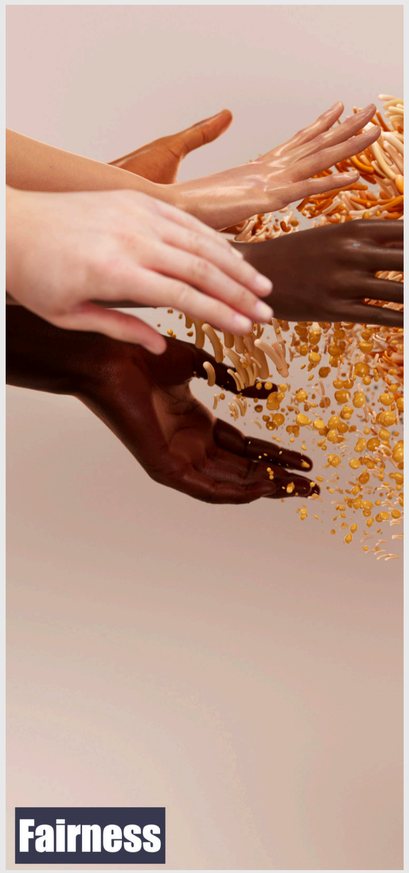
- Under-approximations of transitive closure
- Combine under- and over- approximations for hyperproperties
- Asynchronous Hyperproperties

Research directions from the talk. The bibliography includes a separately identified 2026 preprint on HyperLTL certificates.

[C] Murali, Trivedi & Zamani, Closure Certificates, HSCC 2024 · [X] Murali, Falah, Trivedi & Zamani, HyperCertificates, 2026 preprint



Safe and Secure CPS



Fairness Specifications



Legal Specifications



Fairness Specifications



Saeid Tizpaz-Niari
UT El Paso



Verya Monjezi
UT El Paso



Gang Tan
Penn State University

Verya Monjezi, Ashutosh Trivedi, Gang Tan, Saeid Tizpaz-Niari:
Information-Theoretic Testing and Debugging of Fairness Defects in Deep Neural Networks. ICSE 2023: 1571-1582

Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, Ashutosh Trivedi:
Fairness-aware Configuration of Machine Learning Libraries. ICSE 2022: 909-920

Verya Monjezi, Ashish Kumar, Gang Tan, Ashutosh Trivedi, Saeid Tizpaz-Niari:
Causal Graph Fuzzing for Fair ML Sofware Development. ICSE Companion 2024: 402-403

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023



Fairness

WISCONSIN.GOV

State of Wisconsin
Department of Corrections

COMPAS

DOC uses the Correctional Offender Management Profiling for Alternative Sanctions tool, commonly known as COMPAS, for criminogenic
assessments and unified case planning. This actuarial risk assessment system contains offender information specifically designed to dete
and inform dynamic case plans that will guide the offender throughout his or her lifecycle in the criminal justice system.

PRO PUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Two Petty Theft Arrests

VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8

Two Drug Possession Arrests

DYLAN FUGETT	BERNARD PARKER
Prior Offense 1 attempted burglary	Prior Offense 1 resisting arrest without violence
Subsequent Offenses 3 drug possessions	Subsequent Offenses None
LOW RISK 3	HIGH RISK 10

ProPublica's main charge:
Black defendant face higher false positive rates

Individuals differing only in protected features, receive equivalent outcomes.

$$\forall p_1, p_2 \text{ and } \forall n \text{ we have that } \mathcal{O}(p_1, n) = \mathcal{O}(p_2, n)$$

Historical COMPAS example. The displayed equation is one attribute-invariance requirement, not a complete definition of fairness.

[P] Angwin et al., Machine Bias, ProPublica, May 23, 2016 · [D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023

A debate over fairness semantics

Different measurements

Predictive calibration, false-positive rates, and false-negative rates evaluate different relationships between predictions, labels, and groups.

Explicit specification

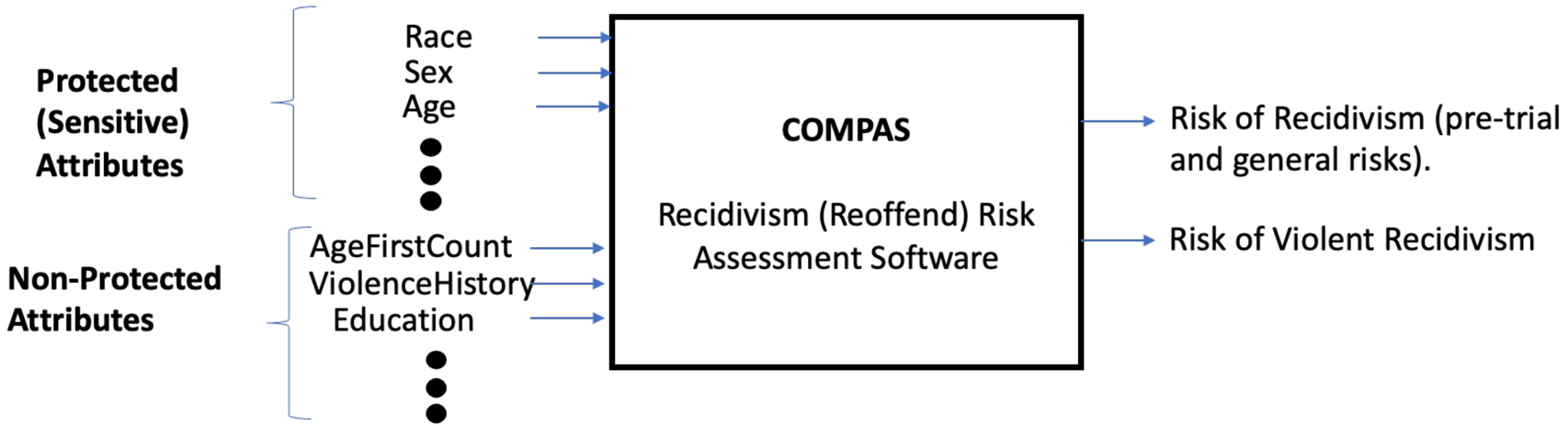
A fairness assessment must name the metric, population, threshold, and observed outcome used as the reference label.

Interpretation

A metric violation identifies a mismatch with that specification. Its explanation and implications require additional evidence about data and deployment.

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023

How to define fairness?



Schematic testing model: attributes available to an auditor need not be direct inputs used by the deployed predictor.

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023

Individual fairness

Fairness through unawareness and awareness

Omitting protected attributes does not remove information carried by correlated features. Similarity-based fairness instead requires a justified input metric and an output relation.

Attribute invariance and causal fairness

Changing a protected attribute while holding all other inputs fixed is a controlled test. A causal counterfactual additionally requires a causal model and a decision about which descendants change.

Quantitative individual fairness

DICE uses information-theoretic quantities to measure violations of its chosen fairness specification, prioritize test cases, and support debugging. These quantities are not universal measures of fairness.

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023

Group fairness: explicit metrics

Outcome rates

Let $r_a = \Pr(\hat{Y} = 1 \mid A = a)$ for group a. Statistical parity difference is $r_a - r_b$. Disparate impact compares r_a / r_b when $r_b > 0$.

Error rates

Equal opportunity difference compares true-positive rates: $TPR_a - TPR_b$. Average odds difference is $\frac{1}{2}[(FPR_a - FPR_b) + (TPR_a - TPR_b)]$.

Evaluation assumptions

Specify the group ordering, favorable outcome, decision threshold, and label Y. Report sample sizes and uncertainty. Small average odds difference can conceal cancellation between error-rate gaps.

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023

Fairness Testing: State of the Art

Themis
FSE'17


Fairness Testing:
Testing Software for Discrimination

Sainyam Galhotra Yuriy Brun Alexandra Meliou

AEQUITAS
ASE'18


Automated Directed Fairness Testing

Sakshi Udeshi Pryanshu Arora Sudipta Chattopadhyay

Black-box
FSE'19


Black Box Fairness Testing of Machine Learning Models

Aniya Aggarwal Pranay Lohia Seema Nagar Kuntal Dey Diptikalyan Saha

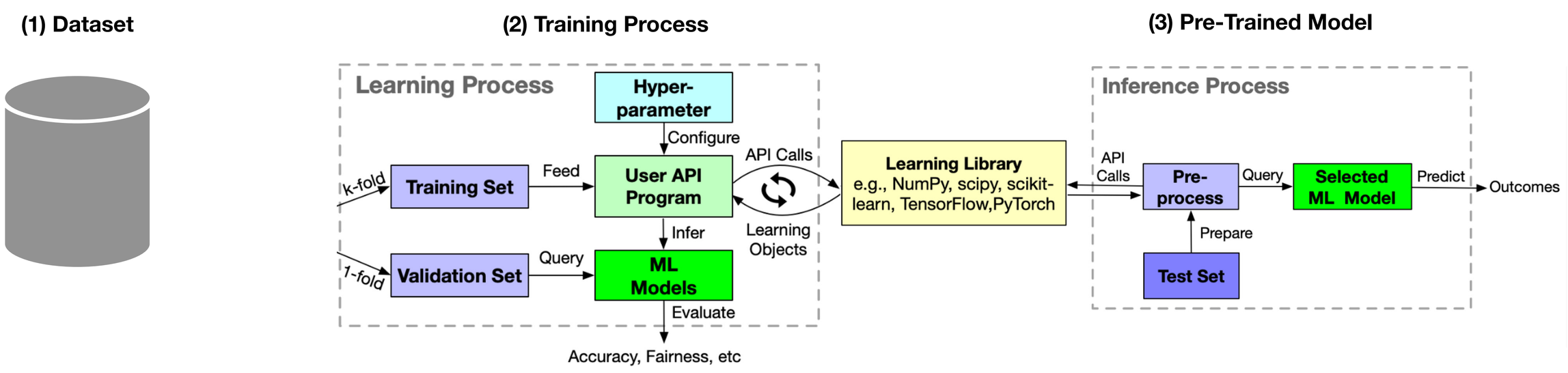
ADT
ICSE'20


White-box Fairness Testing through Adversarial Sampling

Peixin Zhang Jingyi Wang* Jun Sun

Selected prior approaches, not an exhaustive survey. DICE adds quantitative testing and debugging support.

[D] Monjezi, Trivedi, Tan & Tizpaz-Niari, DICE, ICSE 2023



Mitigation for Fairness/Bias Reduction

- Pre-processing
- Post-processing
- In-processing

Adversarial Debiasing AIES'18

Mitigating Unwanted Biases with Adversarial Learning

Brian Hu Zhang Blake Lemoine Margaret Mitchell

Prejudice Remover ICDM'12

Decision Theory for Discrimination-aware Classification

Faisal Kamiran*, Asim Karim†, and Xiangliang Zhang*

Exponentiated Gradient ICML'18

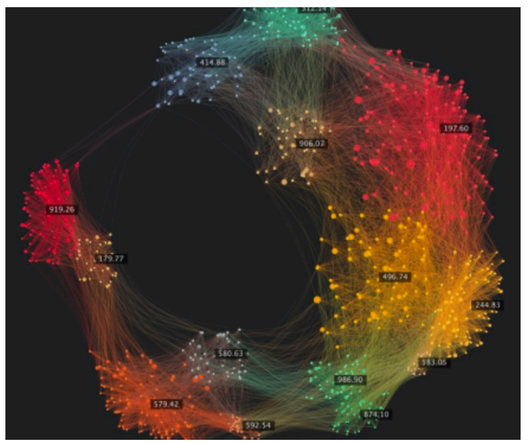
A Reductions Approach to Fair Classification

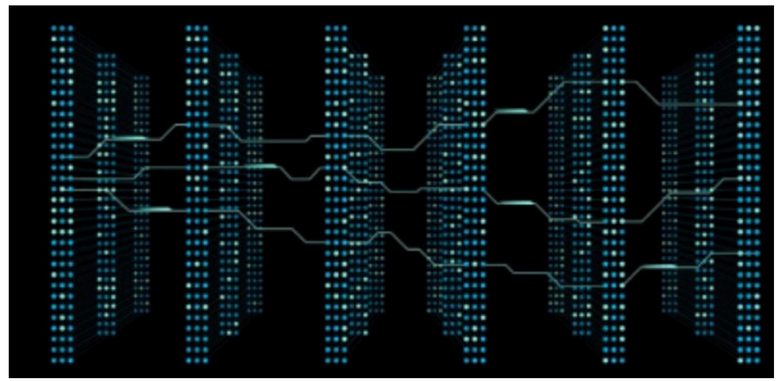
Alekh Agarwal [1] Alina Beygelzimer [2] Miroslav Dudík [1] John Langford [1] Hanna Wallach [1]

These are intervention points. Each intervention must be evaluated against the chosen fairness metric and predictive-performance measure.

[F] Tizpaz-Niari, Kumar, Tan & Trivedi, Parfait-ML, ICSE 2022







How do hyperparameters affect fairness and predictive performance?

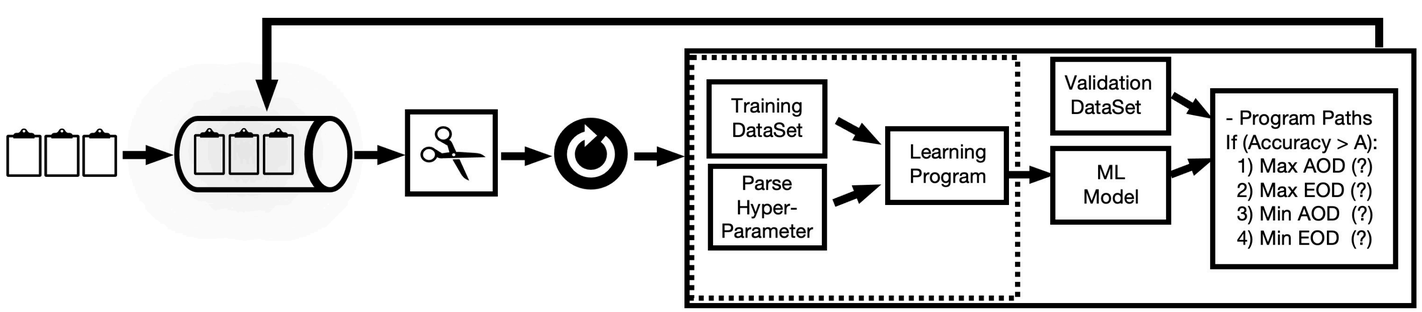
Depth of decision-tree, #layers/neurons in DNN, choice of loss-functions

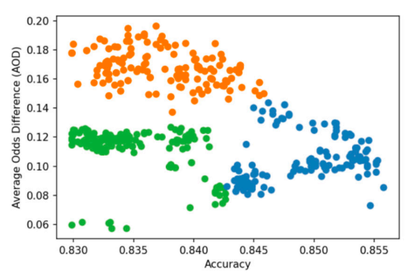
A configuration may improve both measures in a case study. Such improvement is not guaranteed for every dataset or fairness criterion.

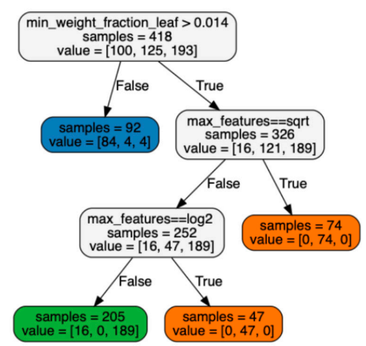
[F] Tizpaz-Niari, Kumar, Tan & Trivedi, Parfait-ML, ICSE 2022

Fairness-aware Configuration of Machine Learning Libraries

https://arxiv.org/pdf/2202.06196.pdf







Challenges:
- Testing individual models
- Modification of datasets and learning algorithms
- Low precision and Scalability.

RQ1: Hyperparameter can aggravate or suppress present biases in the dataset.
RQ2: Mutation-based Evolutionary algorithms are effective search technique.
RQ3: Some hyperparameters systematically introduce biases.
RQ4: Parfait-ML outperforms Exp. Gradients and Fairway mitigation techniques.

Reusable

Available

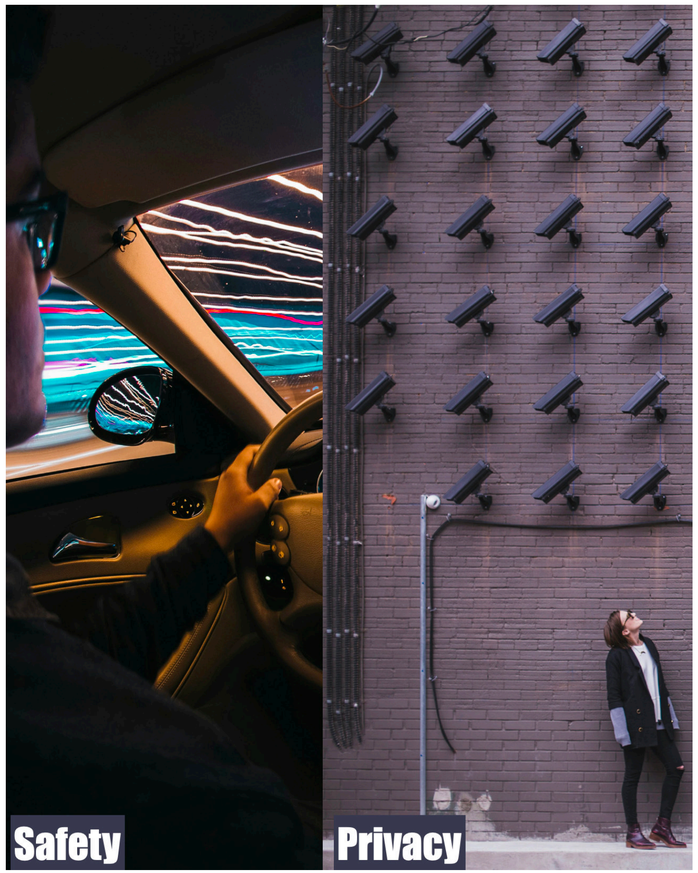
Our **open-source** tool **Parfait-ML** and all experimental subjects are **publicly available**:
https://github.com/Tizpaz/Parfait-ML

Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, Ashutosh Trivedi:
Fairness-aware Configuration of Machine Learning Libraries. ICSE 2022: 909-920

Verya Monjezi, Ashutosh Trivedi, Gang Tan, Saeid Tizpaz-Niari:
Information-Theoretic Testing and Debugging of Fairness Defects in Deep Neural Networks. ICSE 2023: 1571-1582

Parfait-ML explores fairness and precision across configurations. Interpret empirical findings within the study's evaluated tasks.

[F] Tizpaz-Niari, Kumar, Tan & Trivedi, Parfait-ML, ICSE 2022



Safe and Secure CPS



Fairness Specifications



Legal Specifications



Legal Specifications



Morgan Wagner
UT El Paso



Krystia Reed
UT El Paso



Saeid Tizpaz-Niari
UT El Paso



Verya Monjezi
UT El Paso



Siva Darian
CU Boulder

Saeid Tizpaz-Niari, Verya Monjezi, Morgan Wagner, Shiva Darian, Krystia Reed, Ashutosh Trivedi:
Metamorphic Testing and Debugging of Tax Preparation Software. ICSE (SEIS) 2023: 138-149

Dananjay Srinivas, Rohan Das, Saeid Tizpaz-Niari, Ashutosh Trivedi, Maria Leonor Pacheco:
On the Potential and Limitations of Few-Shot In-Context Learning to Generate Metamorphic Specifications for Tax Preparation Software. CoRR abs/2311.11979 (2023)

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023







How to detect bugs in a **legal-critical system** such as tax-preparation software?

- Absence of **oracle**
- Lack of reliable datasets
- Computationally infeasible

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

Legal specifications and metamorphic relations

Motivation from precedent

The analogy with similar legal cases motivates comparing related software inputs. A valid software relation still needs an explicit rule and its conditions.

General form

If R(x, y) holds for two admissible inputs, require S(O(x), O(y)). R specifies the input relation and S specifies the expected output relation.

Testing consequence

A violated relation supplies a candidate counterexample. Confirm the applicable rule and assumptions before classifying it as a software defect.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

A conditional metamorphic test

A simplified deduction model

Hold income, filing status, and all other eligibility conditions fixed. Suppose the only rule change is a larger permitted standard deduction, $d_2 \geq d_1$.

Expected relation

In this model, taxable income is $t(d) = \max(0, I - d)$, so $t(d_2) \leq t(d_1)$. If the tax schedule τ is nondecreasing, then $\tau(t(d_2)) \leq \tau(t(d_1))$.

Boundary of the example

This does not imply that every older taxpayer receives a larger refund. Credits, withholding, phaseouts, and other age-dependent rules can matter. Actual tests must use a specified tax year and complete eligibility conditions.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

Metamorphic testing and debugging of tax software

Method

Derive relations with domain experts, generate related inputs, compare outputs, and inspect counterexamples with explanatory models.

Reported defects

The study found missing eligibility conditions and unexpected behavior near zero-valued returns. These are findings about the studied software versions and test properties.

Student reading

Tizpaz-Niari, Monjezi, Wagner, Darian, Reed, and Trivedi. Metamorphic Testing and Debugging of Tax Preparation Software. ICSE-SEIS 2023. The paper supplies the precise properties and experimental context.

[T] Tizpaz-Niari et al., Tax software testing, ICSE-SEIS 2023

From Legal-Critical to Socio-Critical

“Socio-critical” describes applications whose software errors can materially affect users. The technical focus remains specification and testing.

Auditing benefits screening software

Published case study

Escher and Banovic (CSCW 2020) compared screening decisions with an independently encoded interpretation of eligibility guidelines.

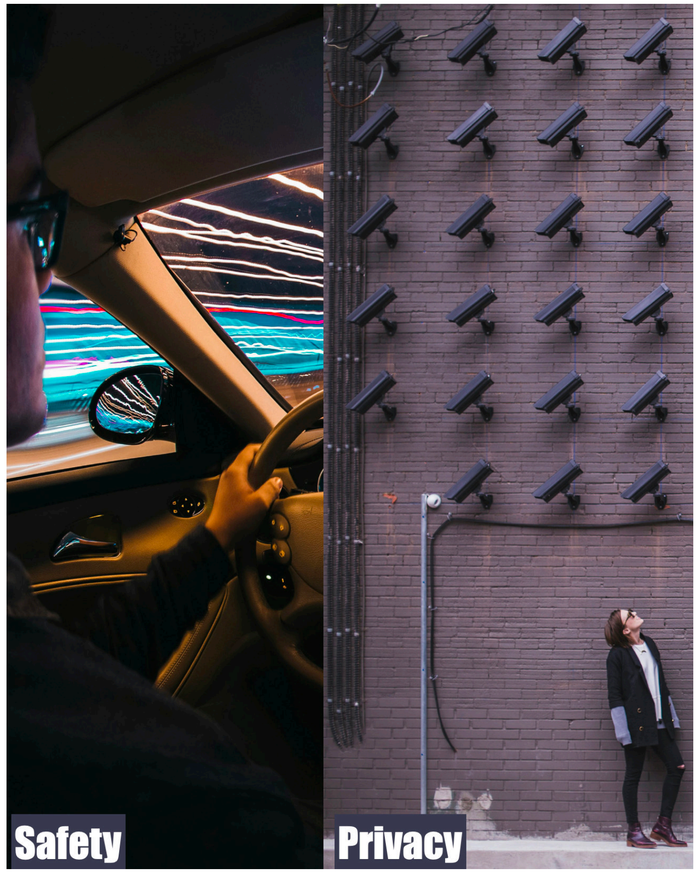
Software-testing lesson

Agreement between an implementation and its specification can be checked using generated inputs. Disagreements require investigation of both the implementation and the reference rules.

Attribution and scope

This is related work by other researchers. The study concerns a benefits screening tool and its historical evaluation, rather than a claim about the present behavior of a deployed system.

[E] Escher & Banovic, CSCW 2020 (related work)







Safe and Secure CPS

Fairness Specifications

Legal Specifications

Privacy, fairness, and legal requirements provide rich problems for specification, verification, and software testing.



CUPLV

Programming Languages and Verification
University of Colorado Boulder

Interested in learning more or doing research on this topic?

Students are welcome to reach out to me at
ashutosh.trivedi@colorado.edu

Email subject: Re. Beyond the Bytes faculty speaker series

ashutoshtrivedi.com · plv.colorado.edu

Reading guide

How to use these slides

Slides 9 and 23 introduce relations between executions. Slides 25–53 develop certificate methods. Slides 55–64 discuss fairness specifications and testing. Slides 66–72 discuss rule-based software.

Definitions and evidence

Formal fairness terms describe chosen mathematical conditions. Historical examples motivate specifications. Neither a headline nor a metric alone supplies a complete causal explanation.

Proofs and tests

A verified certificate establishes its stated property under the model assumptions. Testing finds counterexamples but generally cannot prove their absence.

Links and terminology

Lettered citations beneath slides link to papers. The next pages provide full titles and authors. Published titles and technical terminology remain unchanged.

References: hyperproperties and certificates

[H] Background

Michael R. Clarkson and Fred B. Schneider. **Hyperproperties.** Journal of Computer Security 18(6), 1157–1210, 2010. DOI

[B] Barrier certificates

Mahathi Anand, Vishnu Murali, Ashutosh Trivedi, and Majid Zamani. **Verification of Hyperproperties for Dynamical Systems via Barrier Certificates.** IEEE TAC 69(10), 6920–6934, 2024. DOI · Author preprint (earlier title)

[C] Transition invariants

Vishnu Murali, Ashutosh Trivedi, and Majid Zamani. **Closure Certificates.** HSCC 2024, Article 10. DOI · Open preprint

References: fairness testing and configuration

[D] Testing and debugging

Verya Monjezi, Ashutosh Trivedi, Gang Tan, and Saeid Tizpaz-Niari. **Information-Theoretic Testing and Debugging of Fairness Defects in Deep Neural Networks.** ICSE 2023. Open paper

[F] Hyperparameter search

Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. **Fairness-aware Configuration of Machine Learning Libraries.** ICSE 2022, 909–920. DOI · Open paper

Suggested reading order

Begin with Parfait-ML for configuration search, then DICE for quantitative testing and localization of defects. Each paper makes its own fairness assumptions explicit.

References: metamorphic testing and further reading

[T] Tax-preparation software

Saeid Tizpaz-Niari, Verya Monjezi, Morgan Wagner, Shiva Darian, Krystia Reed, and Ashutosh Trivedi. **Metamorphic Testing and Debugging of Tax Preparation Software.** ICSE-SEIS 2023, 138–149. DOI · Open paper

[X] Related development: preprint

Vishnu Murali, Amin Falah, Ashutosh Trivedi, and Majid Zamani. **HyperCertificates: Verification of Discrete-time Dynamical Systems against HyperLTL Specifications.** arXiv:2605.00752, 2026. Open preprint

Additional perspective

Saeid Tizpaz-Niari, Shiva Darian, and Ashutosh Trivedi. **Metamorphic Debugging for Accountable Software.** arXiv:2409.16140, 2024. Open preprint

References: historical motivation and related work

[P] COMPAS reporting

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. **Machine Bias.** ProPublica, May 23, 2016. Original reporting

[A] Tax audit study

Hadi Elzayn, Evelyn Smith, Thomas Hertz, Arun Ramesh, Robin Fisher, Daniel E. Ho, and Jacob Goldin. **Measuring and Mitigating Racial Disparities in Tax Audits.** SIEPR working paper, January 2023. Study and paper

[E] Benefits screening study

Nel Escher and Nikola Banovic. **Exposing Error in Poverty Management Technology: A Method for Auditing Government Benefits Screening Tools.** PACM HCI 4(CSCW1), Article 64, 2020. DOI

Image acknowledgements

Thank you to the following creators for the images shared through Unsplash and used in this presentation.

Samuele Errico Piccarini
Car interior / safety imagery · View image on Unsplash

Matthew Henry
Surveillance cameras / privacy imagery · View image on Unsplash

Google DeepMind · artwork by Aurora Mititelu
Hands and particles / fairness imagery · View image on Unsplash

Tingey Injury Law Firm
Lady Justice / legal requirements imagery · View image on Unsplash

Martha Dominguez de Gouveia
Clinic waiting area / timing example · View image on Unsplash

Mr Cup / Fabien Barral
Archival photographs / historical data imagery · View image on Unsplash