# Active Formalization in Artificial and Human Reasoners (YIP)

**(FA9550-18-1-0052)**

**PI: John Licato (University of South Florida)**


**AFOSR Program Review:**
**Computational Cognition and Machine Intelligence Program**
**(10/6/20 – 10/8/20, Arlington, VA via ZoomGov)**

# Active Formalization in Human and Artificial Reasoners (John Licato)

## Research Objectives:

- **O1** - Conduct research into reasoning over representational systems.
- **O2** - Investigate the components of active formalization (AF).

## Technical Approach:

**Year 1:** Develop theory of representational systems and identify, define, and collect data for tasks to test

**Year 2:** Develop theory of active formalization; implement algorithms for types of AF on automated reasoner MATR

**Year 3:** Refine algorithms, rigorously test on real-world applications, generalize

## Key Scientific Contributions:

- expand our understanding of the cognitive and logical roots of "good" formal representations in machines and people
- tools and methods for better reasoning (both human and artificial) over formal and informal representational systems

## DoD Benefits:

Advances in automated reasoning, leading to:

- an artificial reasoner being able to reason about, and repair or improve, its own representational systems in order to make it better match recently obtained information.
- an ability to deploy artificial agents with rules of behavior, even when those rules/laws contain informal terminology
- more robustness in automated reasoning about the limitations of, and purpose behind, such rules/laws

# List of Project Goals

1. **O1 -** Conduct research into reasoning over representational systems.
   - Conduct research into frameworks for representational systems, both human and artificial, in order to understand, and ultimately computationally model, reasoning over them.
   - Develop and test a framework for reasoning over both formal and informal representational systems. Explore applications of such a framework to real-world situations that artificial agents (e.g. autonomous robots) might find themselves in.
   - Determine whether such a framework can be expressive enough to apply the types of criteria used by human reasoners to evaluate representational systems.
2. **O2 -** Investigate the components of active formalization.
   - Develop an account of active formalization to the point where it can be modeled computationally.
   - Determine what features an automated reasoner capable of performing active formalization must have, and incorporate as many of these as possible into one (such as the automated reasoning framework MATR3).

# Progress Towards Goals (or New Goals)

**Year Three Goals: Applications of Active Formalization**

- Develop MATR codelets for ADR (**O1**,**O2**)
- Rigorously test real-world applications of active formalization and squeezing algorithms (**O2**)

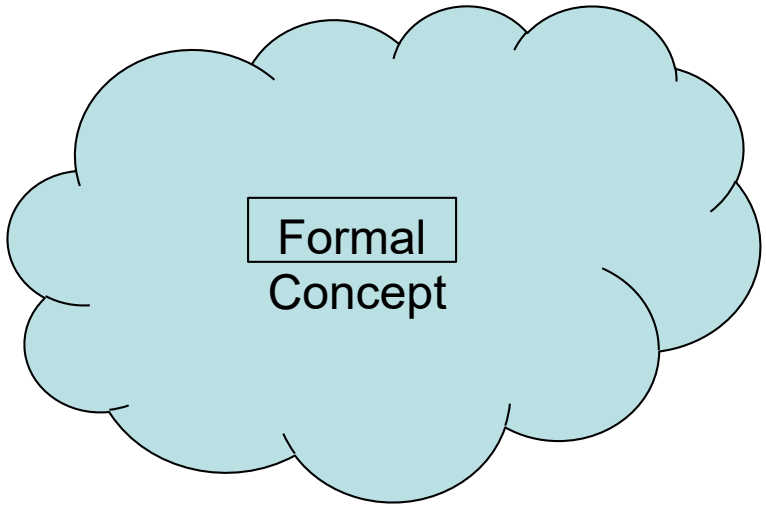
**New Accomplishments (as of Summer 2020)**

- Developed WG-A, a framework for analogical argumentation based on the articulation model
- Gödelian Speedup Theorem
  - Completed metalogical, machine-friendly formalization
  - First to have machine-discovered and machine-verified a Gödelian speedup theorem (in MATR)
- Developing a gamified framework for human-machine agreements

# Active Formalization via Squeezing Algorithms

Advocate

Necessary conditions

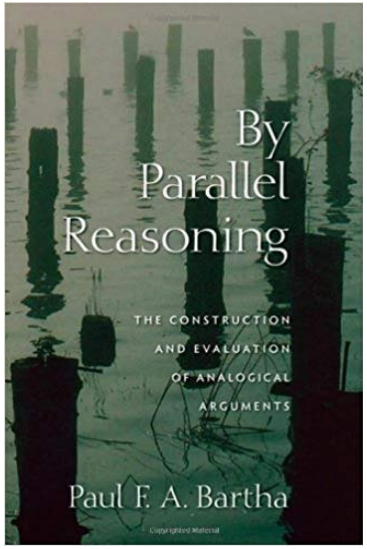



Sufficient conditions

Critic

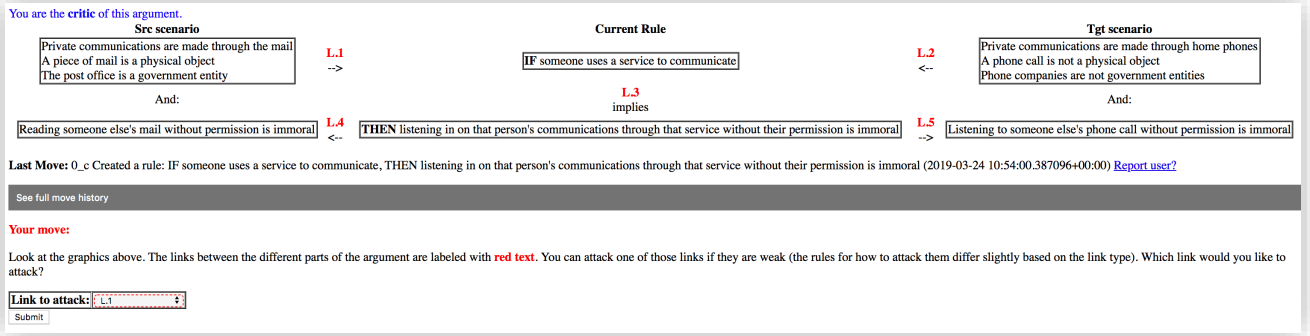
# Bartha's (2010) Articulation Model:

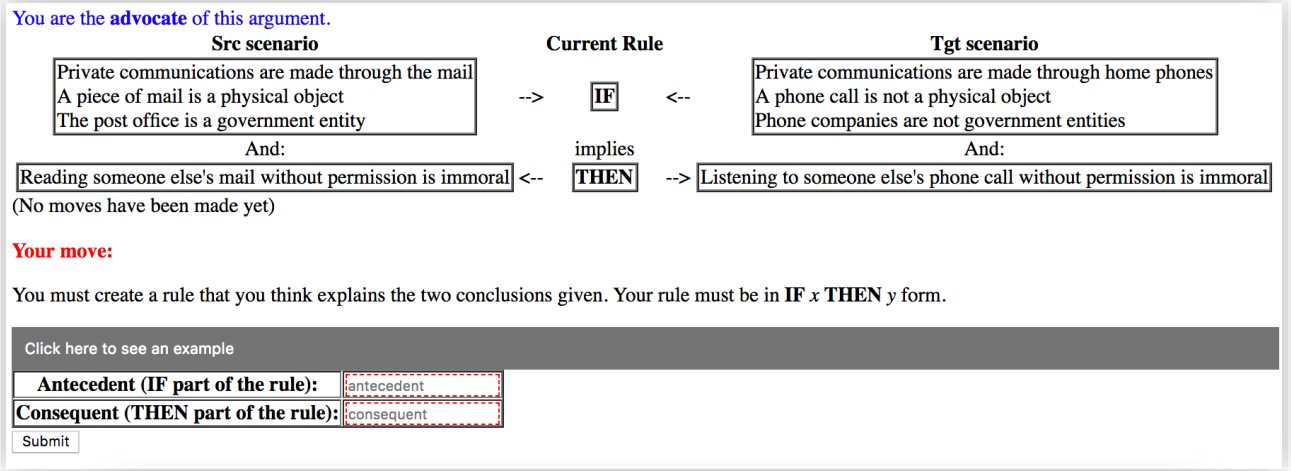
- What is a *good* analogical argument? What makes dialogical moves in assessing an analogical argument *relevant*?
- **Prior Association:** "a clear connection, in the source domain, between the known similarities (the positive analogy) and the further similarity that is projected to hold in the target domain (the hypothetical analogy)." (94)
  - The first goal of an [analogical] argumentative dialogue should be making the PA explicit, as:
    - It allows us to clarify "whether there is reason to think the same kind of connection could obtain in the target domain."
    - We can "classify and evaluate analogical arguments on the basis of [the PA]."
  - "Upper bound thesis" – The PA is an upper bound on the strength of the analogical inference allowed; "at best, we may conclude that an association of similar strength holds in the target domain" (103).
- **Potential for Generalization**: The PA must be transferable to the target domain in such a way that supports the conclusion of the analogical argument.
- We might then say that a move in an argumentative dialogue about analogy is **relevant** iff it contributes to PA or PfG
- Iterative elaborations carried out through interactions between an **advocate** and a **critic**

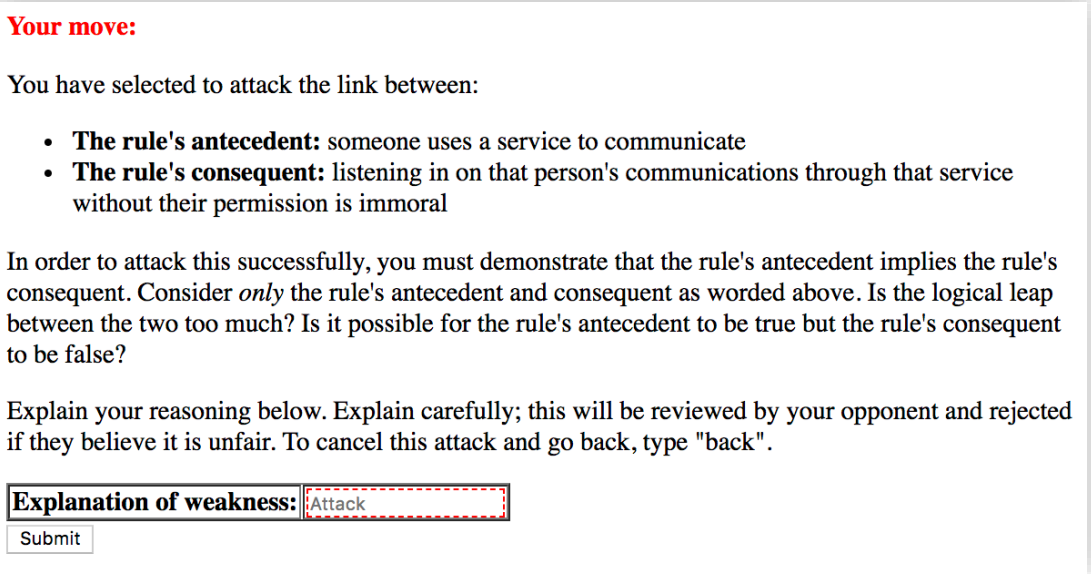


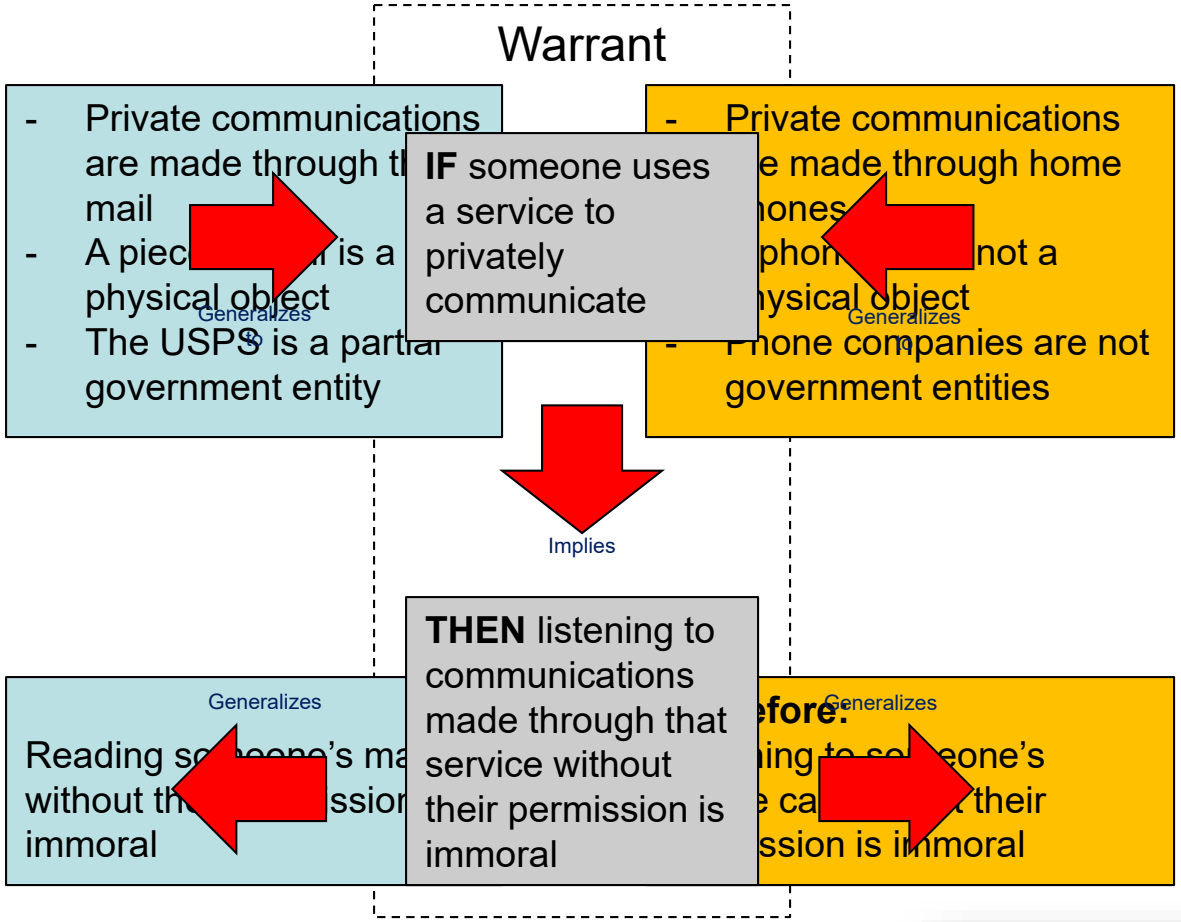
# WG-A

- A “game” to evaluate a single analogical argument
- Variation of the ***w**arrant **g**ame* --- game where shaping the warrant and making it resistant to attacks is the central goal
- Game played through an in-browser app; no other communication between players; allowed moves are highly restricted
- Breaks down complex task of analogical argument assessment into smaller (more AI-reachable) reasoning tasks

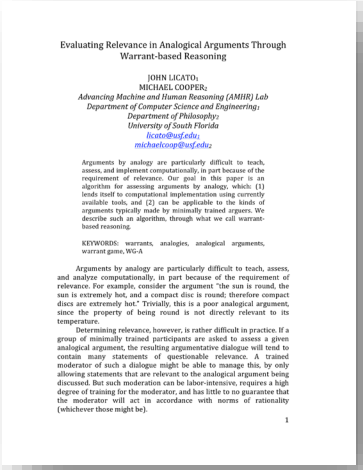





- Positive and negative analogies combined into set of fact pairs / factors. The set of explicit factors are those which are displayed, and can be altered.
- **Warrant**: something like a hybrid between the PA and PfG. Toulmin-esque, but not quite. Currently must be in "if-then" form, and capture relationships from (P ∪ N) to Q and (P* ∪ N*) to Q*
- **Red arrows**: critical links which connect the pieces of the argument together (and make it work).
- **Claim:** Weaknesses in the PA or its PfG can be approximately captured through attacks to one of the critical links.



Evaluating Relevance in Analogical Arguments Through Warrant-based Reasoning

JOHN LICATO[1]
MICHAEL COOPER[2]
*Advancing Machine and Human Reasoning (AMHR) Lab*
*Department of Computer Science and Engineering[1]*
*Department of Philosophy[2]*
*University of South Florida*
licato@usf.edu[1]
michaelcoop@usf.edu[2]

Arguments by analogy are particularly difficult to teach, assess, and implement computationally, in part because of the requirement of relevance. Our goal in this paper is an algorithm for assessing arguments by analogy, which: (1) lends itself to computational implementation using currently available tools, and (2) can be applicable to the kinds of arguments typically made by minimally trained arguers. We describe such an algorithm, through what we call warrant-based reasoning.

KEYWORDS: warrants, analogies, analogical arguments, warrant game, WG-A

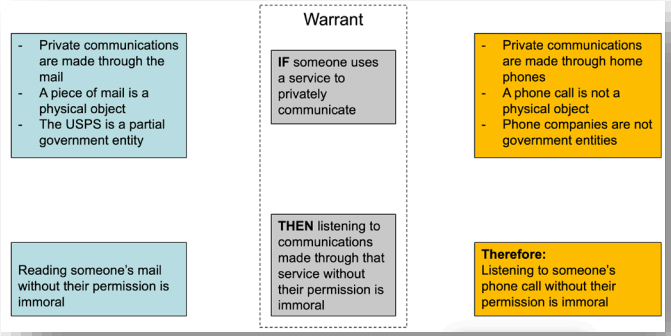
Arguments by analogy are particularly difficult to teach, assess, and analyze computationally, in part because of the requirement of relevance. For example, consider the argument "the sun is round, the sun is extremely hot, and a compact disc is round; therefore compact discs are extremely hot." Trivially, this is a poor analogical argument, since the property of being round is not directly relevant to its temperature.

Determining relevance, however, is rather difficult in practice. If a group of minimally trained participants are asked to assess a given analogical argument, the resulting argumentative dialogue will tend to contain many statements of questionable relevance. A trained moderator of such a dialogue might be able to manage this, by only allowing statements that are relevant to the analogical argument being discussed. But such moderation can be labor-intensive, requires a high degree of training for the moderator, and has little to no guarantee that the moderator will act in accordance with norms of rationality (whichever those might be).

1

Licato, J. & Cooper, M. (2019). Evaluating Relevance in Analogical Arguments through Warrant-based Reasoning. In Proceedings of the European Conference on Argumentation (ECA 2019).

WG-A ensures relevance in analogical argumentation

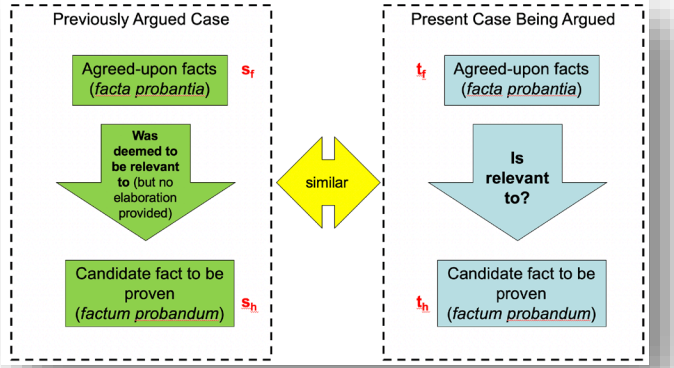




Licato, J. & Cooper, M. (2019). Evaluating Relevance in Analogical Arguments through Warrant-based Reasoning. In Proceedings of the European Conference on Argumentation (ECA 2019).

---

WG-A can be used for extracting evidentiary hypotheses in legal reasoning

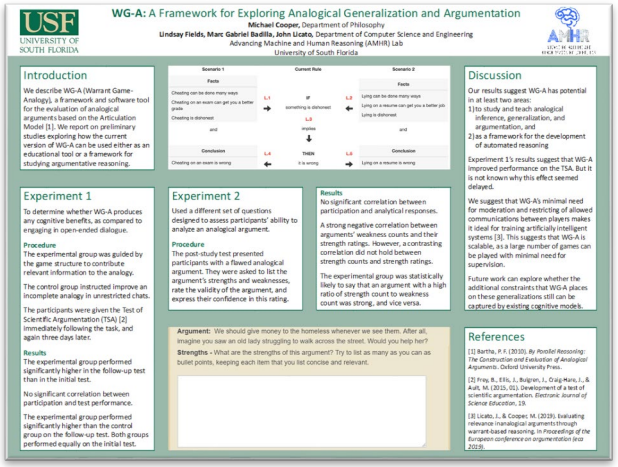




Licato, J. & Cooper, M. (2020). Assessing Evidence Relevance By Disallowing Direct Assessment. *In Proceedings of the 12th Conference of the Ontario Society for the Study of Argumentation.*

---

WG-A improves short-term performance on tests of scientific reasoning (compared to unrestricted argumentative dialogues)





Cooper, M., Fields, L., Badilla, M., & Licato, J. (2020). WG-A: A Framework for Exploring Analogical Generalization and Argumentation. In *Proceedings of the 42nd Cognitive Science Society Conference (CogSci 2020).*

# Do WG-A and other argumentation games have value as tools of introspection and implicit bias inoculation?

## USF selects 23 research projects for funding in anti-racism effort

SEPTEMBER 9, 2020 | RESEARCH AND INNOVATION

A University of South Florida research task force working to address racial issues and attitudes on a local, national and global scale has selected 23 projects exploring a wide range of issues in systemic inequality, economic and health disparities, Black history and contemporary challenges for funding.

The USF Research Task Force on Understanding and Addressing Blackness and Anti-Black Racism in our Local, National and International Communities, which was **first announced** by the university in July, selected the projects as a first-of-its-kind initiative designed to create deeper understanding of complex issues while forging solutions and productive community partnerships. The effort was prompted by several factors, including the long-standing issues of racism and institutional violence brought to the forefront by the recent deaths of Black men, women and children due to excessive force from law enforcement, the disproportionate impact of COVID-19 on the nation's Black communities and other concerns.

Projects spanning eight USF colleges and all three campuses in Tampa, St. Petersburg and Sarasota-Manatee will be part of the year-

Lindsay Fields

Dr. David Ponton

**Argumentation Games to Cognitively Inoculate Against Anti-Black Bias.**

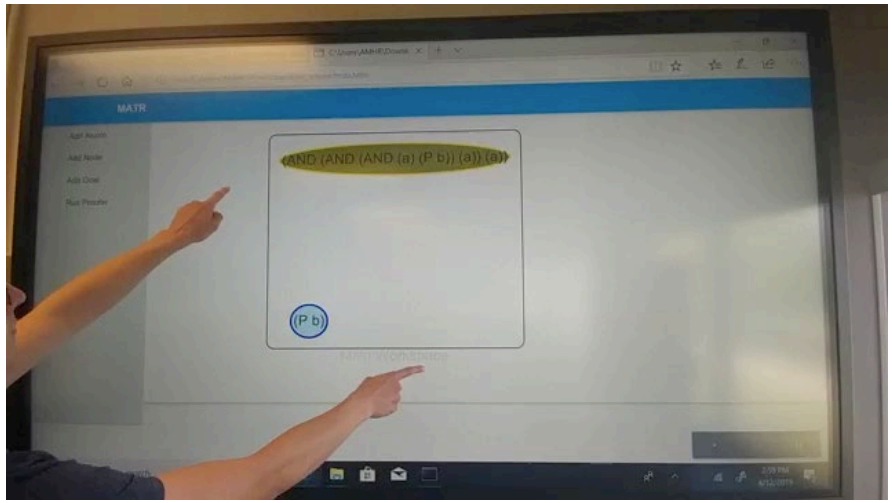
**PI:** John Licato, College of Engineering

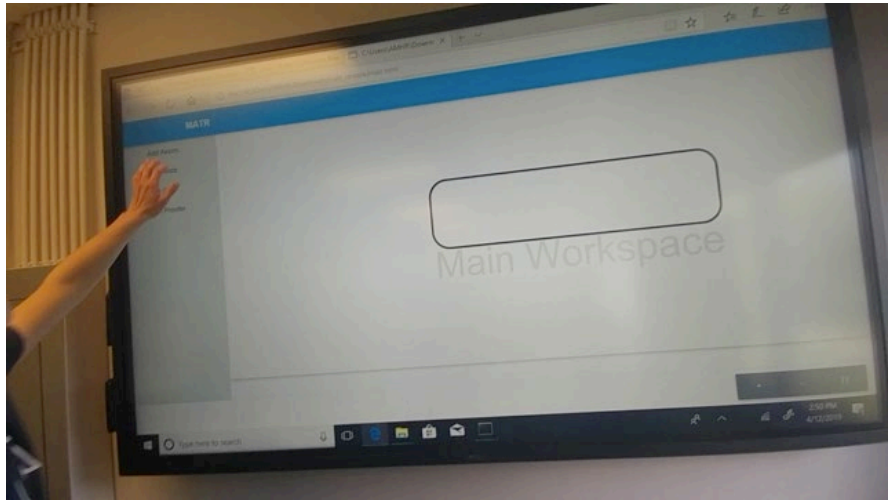
**Community Partner:** Oakridge High School, Orlando

The project will study structured argumentation games (SAGs) as a means of inoculating against anti-Black racism. SAGs are dialogue-based games in which the interactions between participants are subject to highly controlled rules designed by artificially intelligent algorithms that can mitigate some of the damaging effects of unrestricted argumentative dialogues, such as what might occur on social media platforms. Similar games have demonstrated cognitive inoculation effects, whereby participants build resistance against misinformation. The project will explore whether SAGs have the potential to effectively reduce anti-Black bias.

# Formal Reasoning with MATR

- Automated theorem proving framework
- Emphases: rapid deployment of new logics, without having to re-write an ATP from scratch every time
- Flexible “codelets” allow us to quickly implement reasoner variants
- Developed in part with support from AFOSR grant w/RPI
- Allows smooth integration of formal / informal / deductive / inductive reasoning



# Remarkability of Speedup

For **any** p.r. function *f* and (sufficiently powerful) proof system T, there is a wff φ such that:

- T proves φ
- T cannot prove φ in less than *f*(⌜ φ ⌝) steps, but
- There is a higher-order proof system which subsumes T, which can prove φ in less than ⌜ φ ⌝ steps!

- Consider a p.r. function that grows ridiculously fast, e.g.

$$f(n) = n^{n^{n^{\cdots^{n}}}} \quad \text{(n times)}$$

- For any proof theory T (subsuming PA), there's a formula $\phi$ that can't be proven by T in less than $f(\ulcorner\phi\urcorner)$ steps
- But by increasing the expressive power of T, that formula's proof becomes *dramatically* shorter (on the order of f(n))
- A powerful demonstration of the benefits of higher-order reasoning
- Basic trick: "This formula cannot be proven by $T$ in less than $f(\ulcorner\phi\urcorner)$ steps."

# Step 1: Create metalogical formalism able to express relations between proof systems

**Towards Concise, Machine-discovered Proofs of Gödel's Two Incompleteness Theorems**

**Elijah Malaby, Bradley Dragun, John Licato**
Advancing Machine and Human Reasoning (AMHR) Lab
Department of Computer Science and Engineering
University of South Florida

**Abstract**

There is an increasing interest in applying recent advances in AI to automated reasoning, as it may provide useful heuristics in reasoning over formalisms in first-order, second-order, or even meta-logics. To facilitate this research, we present MATR, a new framework for automated theorem proving explicitly designed to easily adapt to unusual logics or integrate new reasoning processes. MATR is formalism-agnostic, highly modular, and programmer-friendly. We explain the high-level design of MATR as well as some details of its implementation. To demonstrate MATR's utility, we then describe a formalized metalogic suitable for proofs of Gödel's Incompleteness Theorems, and report on our progress using our metalogic in MATR to semi-autonomously generate proofs of both the First and Second Incompleteness Theorems.

**Introduction**

An emerging body of literature seeks to apply the recent advances of machine learning and deep networks to the field of automated theorem proving. For example, given a partially completed deductive proof, deciding which inference rules to apply might be a task that modern AI is particularly well-suited to (Wang et al. 2017; Piotrowski and Urban 2019; Kaliszyk et al. 2018; Lederman, Rabe, and Seshia 2018; Kaliszyk, Chollet, and Szegedy 2017; Alemi et al. 2016). Improved decision-making heuristics in automated reasoning are especially important in proofs using non-classical formalisms, such as second-, higher-, or meta-logics. Such logics can sometimes allow for the expression of complex proofs in far fewer steps than might be required in a first-order logic (Buss 1994; Smith 2007). However, this increased expressive power also considerably expands the search space of any proof done in such logics, mandating the need for said improved heuristics.

However, a platform to easily experiment with applying AI to a plurality of logical formalisms does not exist; at least not in a way that jointly satisfies desiderata that we will state shortly. In this paper, we will describe our progress in addressing these goals by presenting MATR (**M**achine **A**ccountability through **T**raceable **R**easoning), a new automated reasoning framework. As a proof-of-concept, we introduce a metalogic capable of expressing proofs of Gödel's Incompleteness Theorems, and show how MATR can be used as a platform for developing AI systems capable of discovering and reasoning over such proofs.



MATR is based on the following design principles:

**P1. The underlying control system should be as formalism-agnostic as possible.** MATR began as an in-house tool to very quickly test the formal representations and inference rules related to variants of the Cognitive Event Calculus (Arkoudas and Bringsjord 2009; Bringsjord and Govindarajulu 2013; Licato et al. 2014; Bringsjord et al. 2014; 2015), whose visual style was inspired by the diagrammatic, flowchart-like aesthetic of Slate (Bringsjord et al. 2008) and the indented subproofs of Fitch-style natural deduction (Barker-Plummer, Barwise, and Etchemendy 2011). Instead of creating an automated theorem prover from scratch for each new formalism, it was decided that a more flexible framework with easily interchangeable parts would be a better long-term strategy.

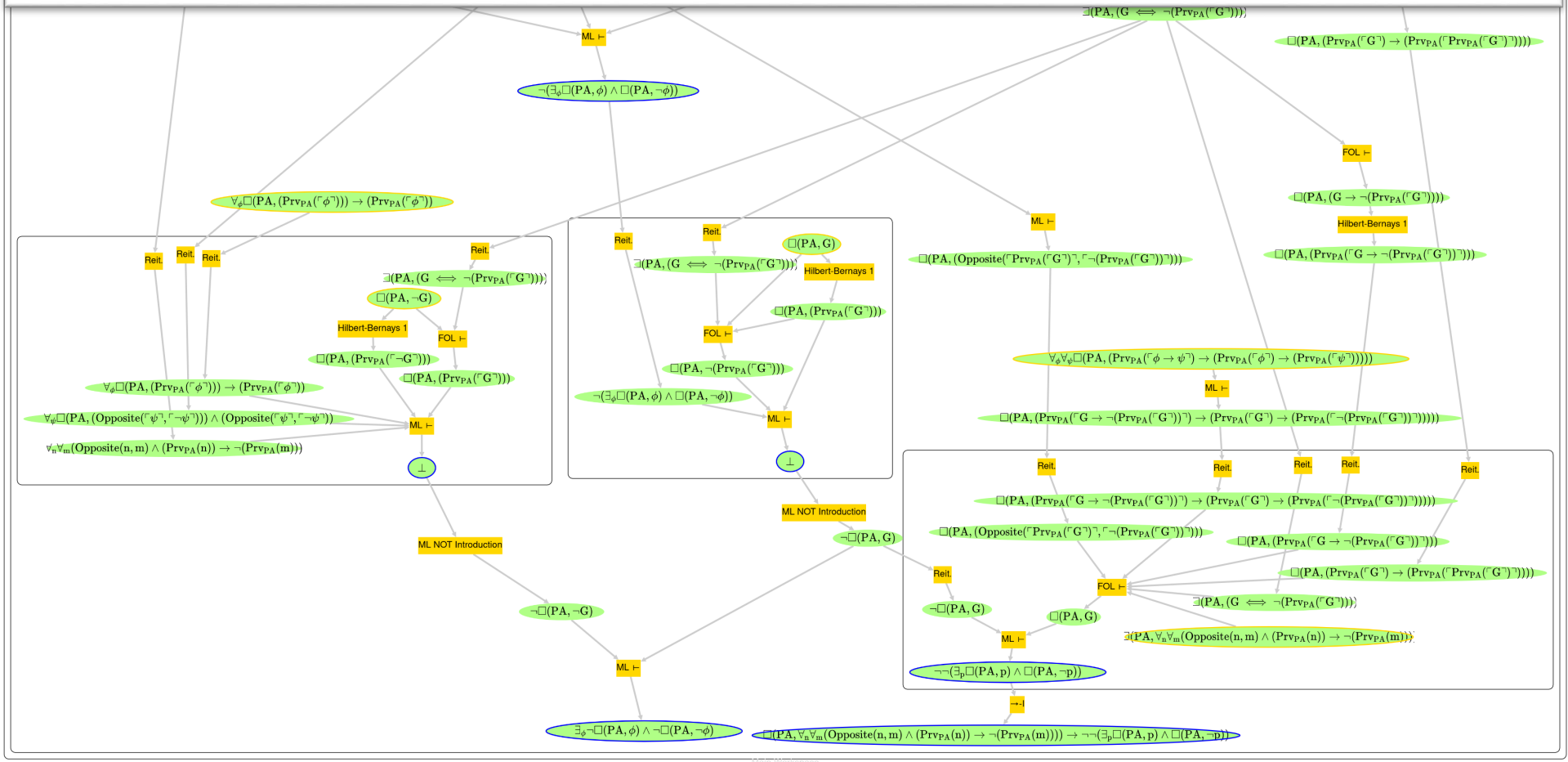
**P2. Semantics should be contained in the codelets and other interchangeable parts.** Any actions requiring semantic understanding of the contents of the nodes in MATR should be contained in one of MATR's interchangeable parts, preferably its codelets. *Codelets* are independently operating programs which perform the bulk of the work in MATR, and are described more later in this section. Syntax checking, carrying out inference rules, recording type information, and even knowing whether a proof is completed are tasks delegated to individual codelets. This is also meant to enable rapid deployment and testing of nontraditional logics (e.g. higher-order, modal, inductive, informal, etc.). One trade-off of this flexibility is that it is entirely possible for a set of codelets to be mutually incompatible. Accordingly, MATR also allows for pre-built *configurations* to be loaded in the form of a YAML text file. For example, if one wishes to use MATR as a natural deduction reasoner for standard first-order Peano Arithmetic, such a configuration will already be available to load.

**P3. Codelets must be programmer-friendly, allowing for easy implementation and changes of inference rules**

Malaby, E., Dragun, B., & Licato, J. (2020). Towards Concise, Machine-discovered Proofs of Gödel's Two Incompleteness Theorems. In *Proceedings of The 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS-33)*. AAAI Press.

$$\exists_{\phi} \neg \Box(\mathrm{PA}, \phi) \wedge \neg \Box(\mathrm{PA}, \neg \phi)$$

$$\Box(\mathrm{PA}, \forall_{\mathrm{n}} \forall_{\mathrm{m}} (\mathrm{Opposite}(\mathrm{n}, \mathrm{m}) \wedge (\mathrm{Prv}_{\mathrm{PA}}(\mathrm{n})) \rightarrow \neg(\mathrm{Prv}_{\mathrm{PA}}(\mathrm{m})))) \rightarrow \neg\neg(\exists_{\mathrm{p}} \Box(\mathrm{PA}, \mathrm{p}) \wedge \Box(\mathrm{PA}, \neg \mathrm{p}))$$



# Step 2: Prove a variant of speedup in this new metalogical formalism

| | | |
|---|---|---|
| 1. | $\forall_x(\phi(x) \leftrightarrow \neg \mathrm{PB}(\mathbb{Z}_i, x, \ulcorner \phi(x) \urcorner))$ | |
| 2. | $\forall_x \forall_y(\mathrm{pc}(\mathbb{Z}_i, x) \wedge \mathrm{gt}(x, y) \rightarrow \mathrm{pc}(\mathbb{Z}_i, y))$ | |
| 3. | $\forall_x \forall_y(\mathrm{gt}(y, x) \wedge \mathrm{PB}(\mathbb{Z}_i, x, \ulcorner \mathrm{pc}(\mathbb{Z}_i, y) \urcorner) \rightarrow \mathrm{pc}(\mathbb{Z}_i, y))$ | |
| 4. | $\mathrm{gt}(\mathsf{f}(+(\mathsf{k}, \mathsf{c})), \mathsf{f}(\mathsf{k}))$ | |
| 5. | $\mathrm{gt}(\mathsf{f}(+(\mathsf{k}, \mathsf{c})), +(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | |
| 6. | $\mathrm{gt}(+(\mathsf{f}(\mathsf{k}), \mathsf{d}), \mathrm{ps}(\ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner))$ | |
| 7. | $\phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \leftrightarrow \neg \mathrm{PB}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}), \ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner)$ | $\forall\mathbf{E}$ : 1 |
| 8. | $\neg \mathrm{PB}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}), \ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner) \rightarrow \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | $\leftrightarrow \mathbf{E}$ : 7 |
| 9. | $\phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \rightarrow \neg \mathrm{PB}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}), \ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner)$ | $\leftrightarrow \mathbf{E}$ : 7 |
| 10. | $\forall_y(\mathrm{gt}(y, \mathsf{f}(\mathsf{k})) \wedge \mathrm{PB}(\mathbb{Z}_i, \mathsf{f}(\mathsf{k}), \ulcorner \mathrm{pc}(\mathbb{Z}_i, y) \urcorner) \rightarrow \mathrm{pc}(\mathbb{Z}_i, y))$ | $\forall\mathbf{E}$ : 3 |
| 11. | $\mathrm{gt}(\mathsf{f}(+(\mathsf{k}, \mathsf{c})), \mathsf{f}(\mathsf{k})) \wedge \mathrm{PB}(\mathbb{Z}_i, \mathsf{f}(\mathsf{k}), \ulcorner \mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \urcorner) \rightarrow \mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c})))$ | $\forall\mathbf{E}$ : 10 |
| 12. | $\forall_y(\mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \wedge \mathrm{gt}(\mathsf{f}(+(\mathsf{k}, \mathsf{c})), y) \rightarrow \mathrm{pc}(\mathbb{Z}_i, y))$ | $\forall\mathbf{E}$ : 2 |
| 13. | $\mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \wedge \mathrm{gt}(\mathsf{f}(+(\mathsf{k}, \mathsf{c})), +(\mathsf{f}(\mathsf{k}), \mathsf{d})) \rightarrow \mathrm{pc}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | $\forall\mathbf{E}$ : 12 |
| 14. | $\mathrm{PB}(\mathbb{Z}_i, \mathsf{f}(\mathsf{k}), \ulcorner \mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \urcorner)$ | |
| 15. | $\mathrm{gt}(+(\mathsf{f}(\mathsf{k}), \mathsf{d}), \mathrm{ps}(\ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner))$ | **Reit** : 6 |
| 16. | $\neg \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | |
| 17. | $\mathrm{PB}(\mathbb{Z}_i, \mathsf{f}(\mathsf{k}), \ulcorner \mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \urcorner)$ | **Reit** : 14 |
| 18. | $\mathrm{gt}(\mathsf{f}(+(\mathsf{k}, \mathsf{c})), \mathsf{f}(\mathsf{k})) \wedge \mathrm{PB}(\mathbb{Z}_i, \mathsf{f}(\mathsf{k}), \ulcorner \mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \urcorner)$ | $\wedge\mathbf{I}$ : 4,17 |
| 19. | $\mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c})))$ | $\rightarrow \mathbf{E}$ : 11,18 |
| 20. | $\mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \wedge \mathrm{gt}(\mathsf{f}(+(\mathsf{k}, \mathsf{c})), +(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | $\wedge\mathbf{I}$ : 5,19 |
| 21. | $\mathrm{pc}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | $\rightarrow \mathbf{E}$ : 13,20 |
| 22. | $\neg\neg \mathrm{PB}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}), \ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner)$ | $\rightarrow \mathbf{E}$ : 8,16 |
| 23. | $\mathrm{PB}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}), \ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner)$ | $\neg\mathbf{E}$ : 22 |
| 24. | $\phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | **Par.Con.** : 21,23 |
| 25. | $\bot$ | $\bot\mathbf{E}$ : 16,24 |
| 26. | $\phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d}))$ | $\bot\mathbf{I}$ : 25 |
| 27. | $\neg \mathrm{PB}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}), \ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner)$ | $\rightarrow \mathbf{E}$ : 9,26 |
| 28. | $\mathrm{PB}(\mathbb{Z}_i, +(\mathsf{f}(\mathsf{k}), \mathsf{d}), \ulcorner \phi(+(\mathsf{f}(\mathsf{k}), \mathsf{d})) \urcorner)$ | **PBI** : 15,26 |
| 29. | $\bot$ | $\bot\mathbf{E}$ : 27,28 |
| 30. | $\neg \mathrm{PB}(\mathbb{Z}_i, \mathsf{f}(\mathsf{k}), \ulcorner \mathrm{pc}(\mathbb{Z}_i, \mathsf{f}(+(\mathsf{k}, \mathsf{c}))) \urcorner)$ | $\neg\mathbf{I}$ : 29 |

pc(Z,n) = partial consistency: Z is consistent on all formulae <=n in size

PB(Z, b, ⌜ p ⌝) = proof-bounded: the shortest proof of p in Z is <= b in size.

- Entire proof above (**P**) is in $Z_i$, thus can be simulated in $Z_{i+1}$
- Size of **P** is fixed, or linearly increases with size of f's definition, since it reasons *about* consistency of $Z_i$ rather than within $Z_i$
- Thus, a proof of phi exists in $Z_{i+1}$ of size $|\ulcorner \mathbf{P} \urcorner| << \ulcorner \mathbf{P} \urcorner << f(\ulcorner \mathbf{P} \urcorner)$. QED

# Interpretive Arguments: Is *x* an instance of *y*?

If expression E occurs in document **D**, E has a setting of S, and E would fit this setting of S by having interpretation I, then E ought to be interpreted as I (Sartor et al. 2014)

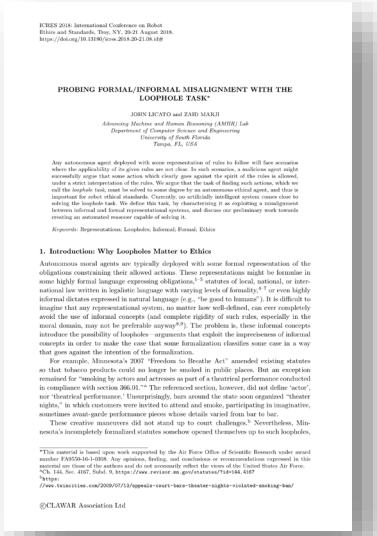
**Some types in legal reasoning (MacCormick and Summers 1991):**

- Arguments from ordinary meaning
- Arguments from technical meaning
- Arguments from precedent
- Arguments from analogy (related: CBR)
- Arguments from history
- Arguments from substantive reasons
- …

- Open-textured predicates: "Exceptions to traffic rules may only be allowed if actions are taken to avert a threat of danger"
- Study of these gives us normative guidelines for active formalization (in both artificial and human reasoners)
- Recognition, evaluation, generation, and use of interpretive argumentation is far beyond current AI SOTA (*yes, even DL!*)

# Loopholes exploit formal/informal misalignment

- An *interpreting method,* in representation $R \in R^*$, takes some description of a case *C* and evidence that *C* is an instance of symbol s, and returns some confidence that *C* is an instance of s.
  - Not necessarily referentially transparent (particularly in informal RSes!)
  - The form of the case and allowed arguments is method-specific (e.g., interpretive arguments vs. fully well-formed formal proofs)
  - A *boolean* interpreting method always returns 'True' or 'False'
  - The determination of which symbol an interpreting method is supposed to recognize is made by the semiotic function
- A representation *recognizes* s *through S* if it has a boolean interpreting method S for s.
  - Note that two methods in different representations may recognize the same symbol, but fail to agree on all possible inputs
  - A representation may also have another method for identifying *exceptions*, which may override S
- Assume (1) the formal representation $F \in F^*$ is supposed to capture an informal representation $I \in I^*$ in order to ban it, and (2) both *F* and *I* recognize s through *F.S* and *I.S*, respectively
  - *F.S* overshoots *I.S* on *c* when *F* returns 'True' for case *c*, but *I* returns 'False'
  - *F.S* undershoots *I.S* on *c* when *F* returns 'False' for case *c,* but *I* returns 'True'

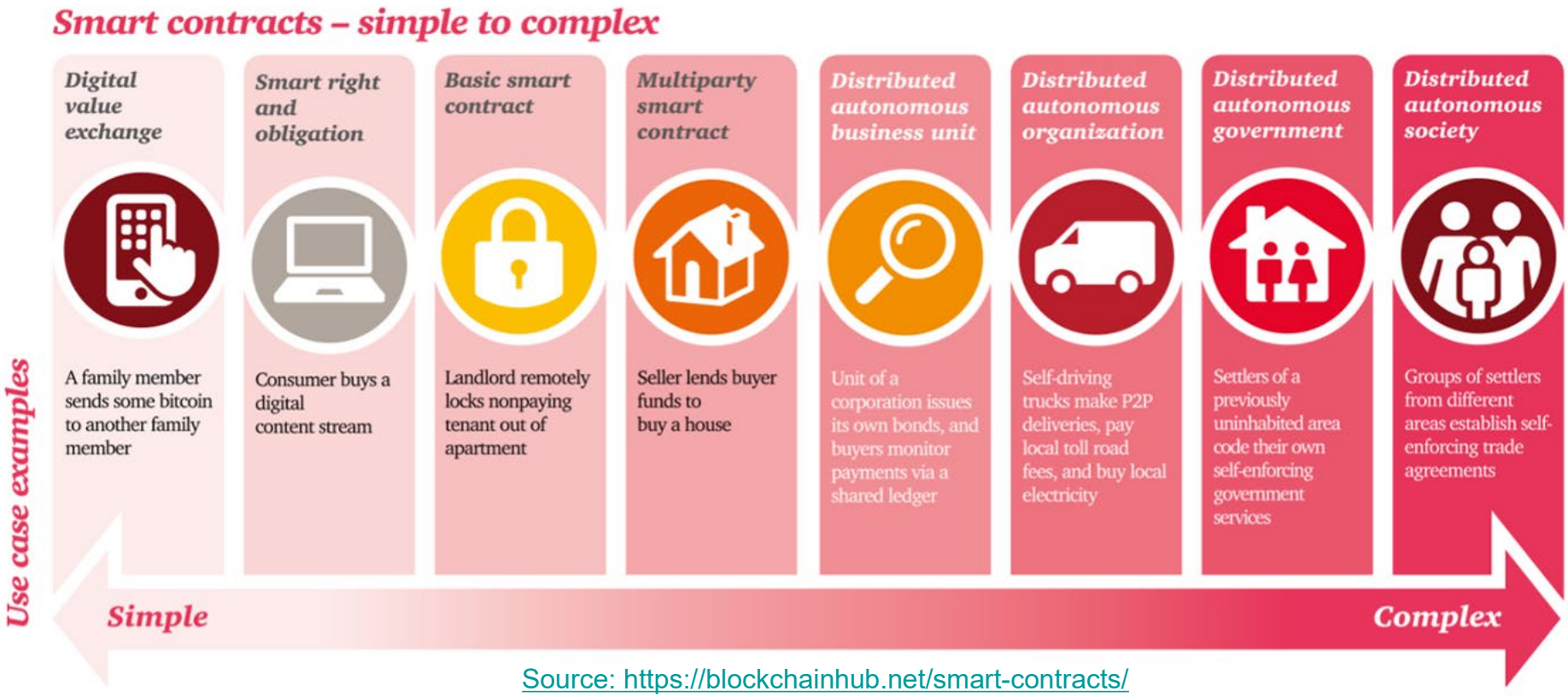



PROBING FORMAL/INFORMAL MISALIGNMENT WITH THE LOOPHOLE TASK*

JOHN LICATO and ZAID MARJI

1. Introduction: Why Loopholes Matter to Ethics

Licato, J. & Marji, Z. (2018). Probing Formal/Informal Misalignment with the Loophole Task. In Proceedings of the 2018 International Conference on Robot Ethics and Standards (ICRES 2018).




Source: https://blockchainhub.net/smart-contracts/

- Contract is specified in code, not natural language (i.e., machine-readable, subject to formal proof methods)
- Participants can assume that if the conditions are met, the contract will independently and reliably trigger (whether or not through blockchain)
- But without open-textured predicates, they will never be a replacement for contracts, laws, rules, etc.

# Claim: Interpretive reasoning is **inevitable** in man/machine agreements, rules, etc. Sooner or later, we need to teach AI to do it better!

> "The interpretation of constitutional principles must not be too literal. We must remember that the machinery of government would not work if it were not allowed a little play in its joints." – *O.W. Holmes*

- Robot is given set of commands:
  - Local / national / international laws
  - Ethical codes of conduct
  - Mission-specific commands
- At some level, these commands will contain informal, open-textured predicates (IOPs); else, they will be inflexible and easy to break using bad-faith antagonist strategy (e.g., fooling DNs)
- How can we sure the robots perform interpretive reasoning / active formalization in a human* way?
- **Solution 1:** Sharpen these IOPs ahead of time
- **Solution 2:** Give the robot the ability to detect, generate, and assess interpretive arguments
- Both solutions are complementary; both also require more research in interpretive argumentation

Problems of Autonomous Agents following Informal, Open-textured Rules*

Ryan Quandt[1] · John Licato[2]
Advancing Machine and Human Reasoning (AMHR) Lab[1,2]
Department of Philosophy[1]
Department of Computer Science and Engineering[2]
University of South Florida

**Abstract**

Autonomous artificial agents—especially those who interact with other intelligent agents—must make use of informal, open-textured rules (IORs), which may express ethical guidelines for behavior, national or international law, or broad moral principles. The design, deployment, and interpretation of IORs, however, is far from a trivial task. Here, we argue that the problems of IORs are unavoidable and should be taken seriously. Drawing from examples in law and statutory interpretation, we describe how the meaning of IORs is constructed through highly context-dependent "interpretive arguments." Lastly, we summarize future work that deals with the problems of IORs and briefly summarize aspects of our current work in this area.

**1 Introduction: Informal, open-textured rules**

The present symposium focuses on the "meaning, value and interdependent effects that AI has on society wherever these machines may interact with humans or other autonomous agents."[1] Whether human beings or artificially intelligent systems of the future, autonomous agents are typically afforded a degree of trust commensurate to the set of rules they are expected to follow, and their ability to faithfully interpret and act in accordance with those rules. For example, a soldier is expected to follow ethical codes, but also to have a reasonable understanding of how those codes are likely to be interpreted. Ideally, the same would be expected of a robot allowed to act autonomously in a similar setting.

Such rules governing acceptable actions are unavoidably context-dependent. As we will argue in this paper, these rules, even when they are seemingly simple, contain elements that are informal and open-textured. This results in two limitations: asymmetry and indeterminacy. We argue that a solution to these problems—adopting a common solution used by people—is to give automated reasoners the ability to properly assess candidate interpretations of given rules by reasoning over their *interpretive arguments* (MacCormick & Summers, 1991; Summers, 2006; Sartor, Walton, Macagno, & Rotolo, 2014; Walton, Sartor, & Macagno, 2016).

Consider a simple example: a robot commanded to `pick up a cup`. In the best of cases, it recognizes a cup, extends its open hand, presses snugly, and lifts. Then it is commanded to do so again, albeit with changes: `pick up the red cup` is the command given (**C**) and there are now three cups on the table: a red, a yellow, and a blue one. It's another success. The trial is set up again with the same command **C**, except there are three cups of varying types on the table with similar reddish hues (Figure 1). An argument could be made that each is a red cup, but which, if any, is the intended object? And assuming the command-giver is not available to clarify, how might an intelligent agent decide which action is most likely correct?

An informal, open-textured rule (IOR) is of this sort: the concepts 'red' and 'cup' in **C** are vague to an extent. 'Red' is on a color spectrum, which can result in ambiguities, and 'cup' may refer to a drinking cup, a trophy cup, a cupping-glass, or a cup that holds writing utensils. But the vagueness is magnified and insolvable with informal, open-textured rules—so we will argue. An IOR occurs when there are concepts that over-shoot or under-shoot (and

*This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-16-1-0308. Any opinions, finding, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

[1] https://sites.google.com/site/aaai19sharedcontext/

1

Quandt, R. & Licato, J. (2019). Problems of Autonomous Agents following Informal, Open-textured Rules. In Proceedings of the AAAI 2019 Spring Symposium on Shared Context.

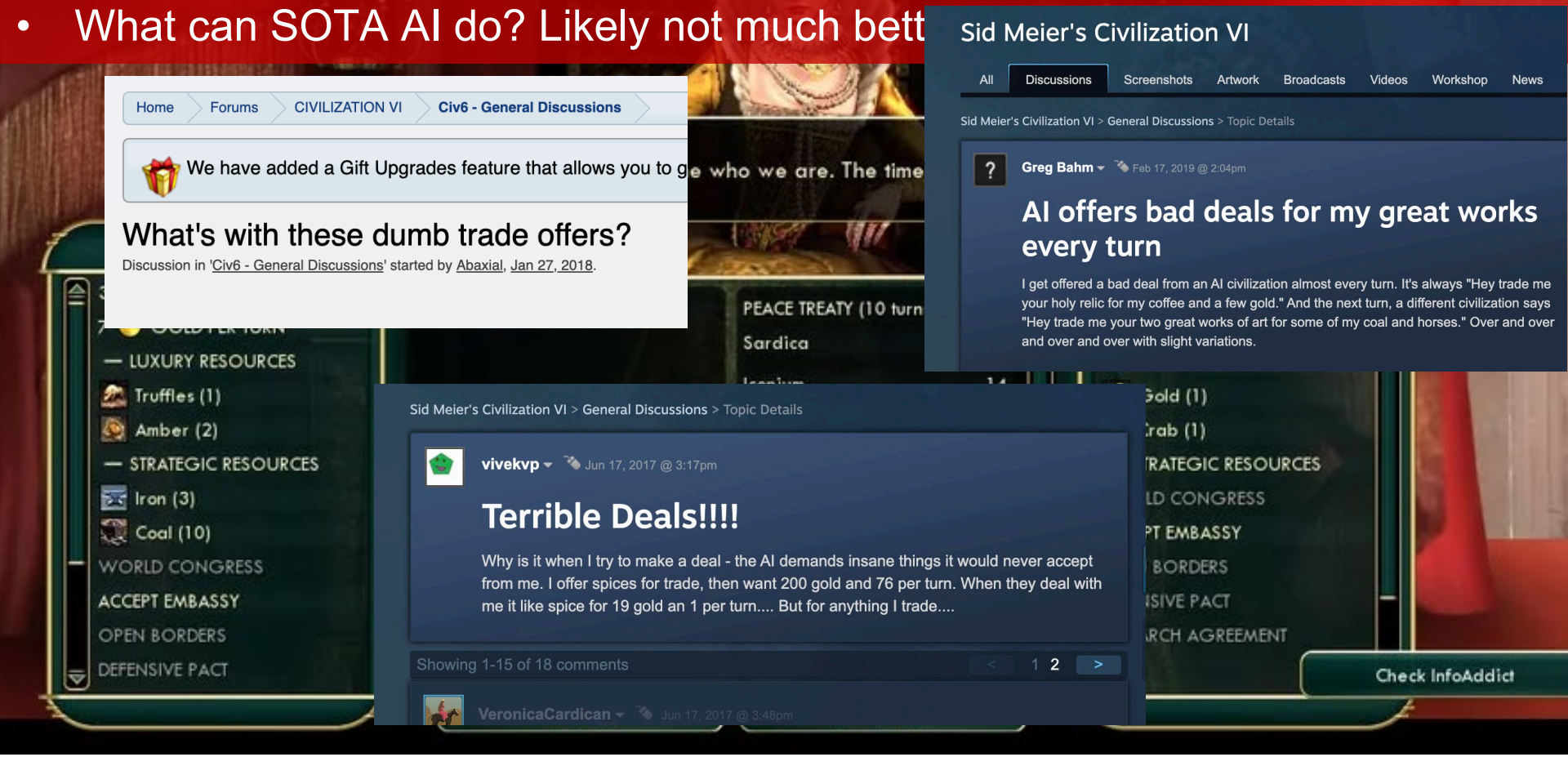
Quandt, R. & Licato, J. (2020). Problems of Autonomous Agents Following Informal, Open-Textured Rules. In Lawless, W.F., Mittu, R., & Sofge, D.A., eds. Human-Machine Shared Contexts. Academic Press.

What would a world look like where people and machines could negotiate, and be held to, smart contracts?

What kind of reasoning is necessary to:

- negotiate smart contracts (with open-texturedness)
- evaluate them (fairness, reasonability, necessity, alignment with long term goals, ...)
- navigate ambiguity in open-textured terms
- generate arguments in support or against interpretations
- evaluate such arguments
- adjudicate conflicts
- etc.

- Strategy games can be an excellent test environment with rich scenarios and contexts
- But agreement mechanisms in most games (all games with AI players) are primitive
- Even within these simplified negotiation domains, the AI reasoning is poor
- What can SOTA AI do? Likely not much bett



# Future Work: *Pactum in Codice*

(working title)

- A framework for human / machine agreements that uses a form of smart contract
- Designed at an abstract level, applicable to many strategy games: Freeciv, Civ 6, CK3, EU4, etc.
- Will allow for study of interpretive reasoning, negotiation-based planning, designing rules for both people and AI to follow, etc.
- Can be used as testbed for models of trust and human-machine teaming

**Activation Conditions**

Time elapsed
Transfer of money
Player presses button
…

Player acts aggressively towards France
Player assists N. Korea in obtaining weapons
…

Combination Logic

**Output Events**

France has a casus belli
Funds are automatically transferred
Ownership of contested resource changes
…

# List of Publications, Awards, Honors, etc. Attributed to the Grant

**Published:**

- Boger, M., Laverghetta Jr., A., Fetisov, N., & Licato, J. (2019). Generating Near and Far Analogies for Educational Applications: Progress and Challenges. In *Proceedings of the 2019 ICMLA Special Session on Machine Learning Applications in Education*.
- Marji, Z., Nighojkar, A., & Licato, J. (2020). Probing the Natural Language Inference Task with Automated Reasoning Tools. In *Proceedings of The 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS-33)*. AAAI Press.
- Malaby, E., Dragun, B., & Licato, J. (2020). Towards Concise, Machine-discovered Proofs of Gödel's Two Incompleteness Theorems. In *Proceedings of The 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS-33)*. AAAI Press.
- Licato, J. (2020). Commentary on Michael Yong-Set, "Getting Down in the MUDs: A Ludological Perspective on Arguers". In *Proceedings of the 12th Conference of the Ontario Society for the Study of Argumentation*.
- Licato, J. & Cooper, M. (2020). Assessing Evidence Relevance By Disallowing Direct Assessment. *In Proceedings of the 12th Conference of the Ontario Society for the Study of Argumentation*.
- Cooper, M., Fields, L., Badilla, M., & Licato, J. (2020). WG-A: A Framework for Exploring Analogical Generalization and Argumentation. In *Proceedings of the 42nd Cognitive Science Society Conference (CogSci 2020)*.
- Quandt, R. & Licato, J. (2020). Problems of Autonomous Agents Following Informal, Open-Textured Rules. In Lawless, W.F., Mittu, R., & Sofge, D.A., eds. *Human-Machine Shared Contexts*. Academic Press.

**Under Review:**

- Fetisov, N. & Licato, J. Improving Automated Essay Scoring Explainability by Using Sentence-Level Scoring. Submitted to *AAAI 2021*.
- Nighojkar, A. & Licato, J. Can Mutual Implication Capture Sentence Equivalence? Submitted to *AAAI 2021*.
- Laverghetta, A., Mirzakhalov, J., & Licato, J. Curriculum Learning for Natural Language Inference. Submitted to *AAAI 2021*.
- Malaby, E. & Licato, J. Revisiting Inductive Proof Automation for Coq. Submitted to *Certified Programs and Proofs (CPP 2021)*.
- Dragun, B., Malaby, E., & Licato, J. A Survey of Conjecture Generation in Artificial Intelligence. Submitted to *Artificial Intelligence Review.*

**Awards and Other Recognition:**

- Licato, J. & Ponton, David. Argumentation Games to Cognitively Inoculate Against Anti-Black Bias. Awarded $29,013 by USF's "Understanding and Addressing Blackness and Anti-Black Racism in Local, National, and International Communities Research" program (9/2020 – 9/2021).