Goal Recognition, Norms, and Principled Disobedience in BDI Agents

Prof. Felipe Meneguzzi

University of Aberdeen, Scotland, UK
felipe.meneguzzi@abdn.ac.uk

RAD-AI @ AAMAS, Cyprus, 2026

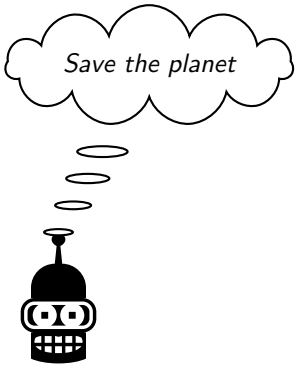
UNIVERSITY OF ABERDEEN

Table of Contents

Motivation

The Compliance Question

- Autonomous agents operate in environments regulated by norms
  - Laws, contracts, social conventions
  - Obligations, permissions, prohibitions
- Classic assumption: agents *should* comply
  - But compliance may be impossible
  - Or may conflict with an agent's goals
  - Or the norm itself may be unjust

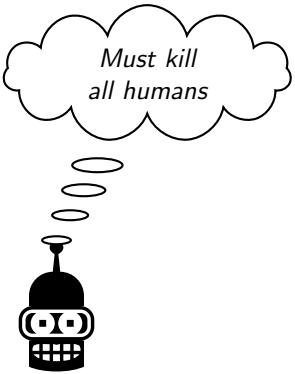


[a] Thomas Arnold, Gordon Briggs, and Matthias Scheutz. "Only Those Who Can Obey Can Disobey: the Intentional Implications of Artificial Agent Disobedience". In: *RAD-AI @ AAMAS*. 2022.

Motivation

The Compliance Question

- Autonomous agents operate in environments regulated by norms
  - Laws, contracts, social conventions
  - Obligations, permissions, prohibitions
- Classic assumption: agents *should* comply
  - But compliance may be impossible
  - Or may conflict with an agent's goals
  - Or the norm itself may be unjust
- **When should an agent disobey?**[a]



[a] Arnold, Briggs, and Scheutz, "Only Those Who Can Obey Can Disobey: the Intentional Implications of Artificial Agent Disobedience".

The Dual Problem

- **Deciding your own behaviour**
  - How does a BDI agent represent and reason about norms?
  - When does goal achievement *require* norm violation?
  - How do we learn what norms apply from observing others?
- **Recognising others' intent**
  - What is another agent trying to achieve?
  - Is their behaviour compliant with norms?
  - How can we detect norm violations from observation?

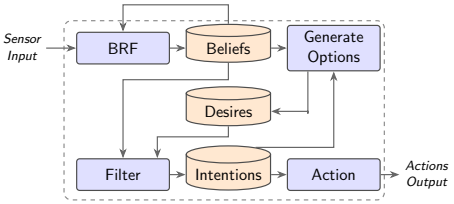
Table of Contents

BDI Agents

A model of *practical* reasoning

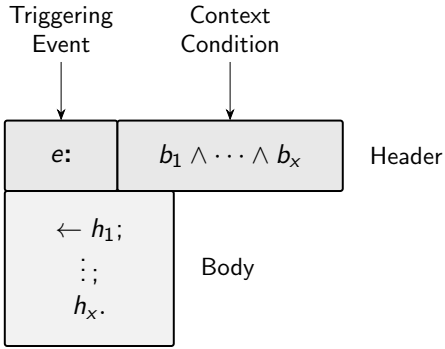
Agent architecture based on three "mental" structures.

- **Beliefs** — agent's knowledge about the world
- **Desires** — goals the agent wants to achieve
- **Intentions** — committed courses of action
- Based on a philosophical model of *practical reasoning*
- Key process: *means-ends reasoning*
  - Typically via a *plan library*
  - Recently: *generalised planning*, *natural language*



BDI Reasoning Cycle

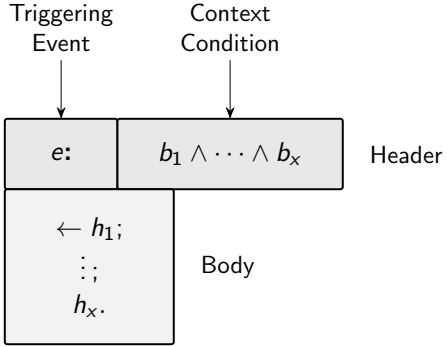
1. Perceive events from the environment
2. Update belief base
3. Select *relevant* plans (triggering event matches)
4. Select *applicable* plans (context condition holds)
5. Add chosen plan to intentions
6. Execute one step of current intentions



Plans in BDI

Plan Library

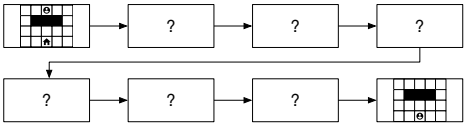
- Plans are *reactive procedures*:
  - Triggering event e
  - Context condition $b_1 \wedge \cdots \wedge b_x$
  - Body: ordered list of actions/sub-goals
- Filters for relevant/applicable plans
- Plan library encodes *designer knowledge*
  - Cannot handle unanticipated situations



Plans in BDI

Plan Library

- Plans are *reactive procedures*:
  - Triggering event e
  - Context condition $b_1 \wedge \cdots \wedge b_x$
  - Body: ordered list of actions/sub-goals
- Filters for relevant/applicable plans
- Plan library encodes *designer knowledge*
  - Cannot handle unanticipated situations



- Plans *focus reasoning*
- Filters for relevant/applicable plans?
- Plan library encodes *designer knowledge*
  - Still limited to what the designer anticipated

BDI Today: Agentic LLMs

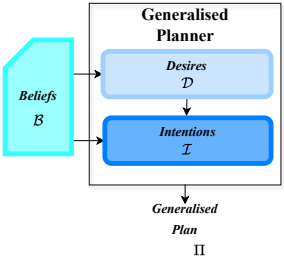
BDI	**Agentic LLM**
Belief base	Context
Plan library	`PLAN.md`
Means-ends reasoning	The LLM itself
Reasoning cycle	Tool-call loop
Triggering event	User prompt

"BDI didn't go away — it just stopped being cited."

Automated Planning in BDI

Beyond Plan Libraries[1]

- Plan libraries are *finite* — they cannot anticipate all situations
- Automated planners generate plans *at runtime*
  - Handle novel situations not foreseen by designers
  - But: planning is computationally expensive
- Recent direction: *generalised planning* in BDI
  - Plans that work across *classes* of problems
  - Bridges plan library robustness with planning flexibility

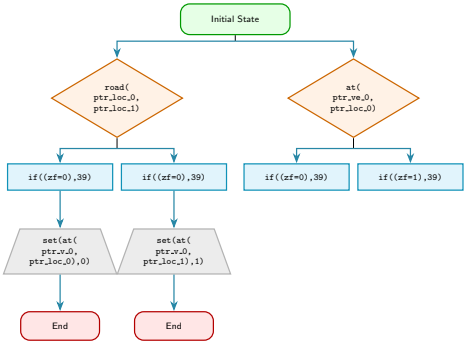


[1] Ramon Fraga Pereira, Nir Oren, and Felipe Meneguzzi. "Generalised BDI Planning". In: *AAMAS*. 2025.

Generalised Planning in BDI

Overview

- Generalised plans: programs with loops and conditionals
- Solve not one problem, but a *class* of related problems
- Key insight for BDI:
  - A BDI agent *is* a generalised planner
  - Reasoning cycle implements a generalised executor
  - Plan library entries are conditional branches
- Formal correspondence enables verification



Generalised Planning in BDI

Overview

- Generalised plans: programs with loops and conditionals
- Solve not one problem, but a *class* of related problems
- Key insight for BDI:
  - A BDI agent *is* a generalised planner
  - Reasoning cycle implements a generalised executor
  - Plan library entries are conditional branches
- Formal correspondence enables verification

```
// Direct move: road exists
+!move(V, L0, L1)
  : road(L0, L1) & at(V, L0)
  ← -at(V, L0);
     +at(V, L1).

// Indirect move: no direct road
+!move(V, L0, L1)
  : not road(L0, L1)
  ← !findRoute(V, L0, L1).
```

Planning for the Environment and Planning for Society

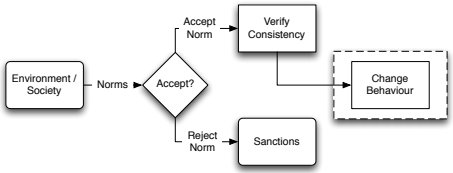
- Plan Synthesis / BDI Programmes (usually):
  - Designed to achieve goals;
  - Cope with environmental restrictions;
  - Explicitly coordinate with other agents.
- The usual assumption is of prior knowledge:
  - Of agent goals;
  - Environmental dynamics;
  - Interaction protocols.
- But how do we deal with dynamic societies and implicit behaviour?
  - Agents with unknown or implicit goals?
  - Dynamic environments?
  - Emerging societal dynamics?

Table of Contents

Norms and Agents

Why we need norms[2]

- Autonomous agents operate in *open* environments;
- Norms regulate these environments;
- They use deontic concepts:
  - **Obligations** — what agents *must* do
  - **Permissions** — what agents *may* do
  - **Prohibitions** — what agents *must not* do
- Challenge: traditional BDI provides *no mechanism* to adapt at runtime to newly accepted norms



[2] Felipe Meneguzzi and Michael Luck. "Norm-based Behaviour Modification in BDI Agents". In: *AAMAS*. 2009.

Norm Representation

- Norms have an operational structure:

$$\langle \nu, \mathit{Act}, \mathit{Exp}, \mathit{id} \rangle$$

  - ν — normative formula
  - Act — activation condition
  - Exp — expiration condition
  - id — norm identifier
- Normative formula: $\mathbf{X}_{\alpha:\rho}\, \varphi \circ \Gamma$
  - $\mathbf{X} \in \{\mathbf{O}, \mathbf{F}\}$ — obligation/prohibition
  - φ — targeted action
  - Γ — constraints on action parameters

LLM guardrail as a norm

"Do not delete files during a read-only task"

- $\nu = \mathbf{F}\ \texttt{delete}(F) \circ \texttt{critical}(F)$
- $\mathit{Act} = \texttt{taskGoal}(\texttt{readOnly})$
- $\mathit{Exp} = \texttt{taskComplete}$

Yet: a tool-call chain

$\texttt{read} \rightarrow \texttt{summarise} \rightarrow \texttt{cleanup} \rightarrow \texttt{delete}$

. . . still wipes critical files: **a norm violation**

Norm Representation

- Norms have an operational structure:

$$\langle \nu,\ Act,\ Exp,\ id \rangle$$

  - ν — normative formula
  - Act — activation condition
  - Exp — expiration condition
  - id — norm identifier
- Normative formula: $\mathbf{X}_{\alpha:\rho}\, \varphi \circ \Gamma$
  - $\mathbf{X} \in \{\mathbf{O}, \mathbf{F}\}$ — obligation/prohibition
  - φ — targeted action
  - Γ — constraints on action parameters

LLM guardrail as a norm

"Do not delete files during a read-only task"

- $\nu = \mathbf{F}\ \texttt{delete}(F) \circ \texttt{critical}(F)$
- $Act = \texttt{taskGoal}(\texttt{readOnly})$
- $Exp = \texttt{taskComplete}$

Yet: a tool-call chain
$\texttt{read} \rightarrow \texttt{summarise} \rightarrow \texttt{cleanup} \rightarrow \texttt{delete}$
. . . still wipes critical files: **a norm violation**

What counts as "critical" is determined by the user's implicit intent.
Only an agent that recognises that intent can disobey the norm **selectively**.

Previous Normative Agent Systems

Two Extremes

- **Blanket plan retractions** (Normative AgentSpeak)
  - Entire plans removed when a norm activates
  - Loses autonomy entirely
- **Every norm checked at every step** (BOID)
  - Computationally wasteful
  - Compliance decision made too early
- Both extremes: *non-compliance is not an option*
- cf. R-HTN[3]: adaptive HTN agents that *modify* plans on violation — more goals reached, fewer violations
- cf. Ethical Rebellion System[4]: dual-process architecture — fast intuitive + deliberative ethical reasoning

[3] Héctor Muñoz-Avila, David W. Aha, and Paola Rizzo. "R-HTN: Rebellious Online HTN Planning". In: *RAD-AI @ AAMAS*. 2025.

[4] Ursula Addison, Othalia Larue, and Matthew Molineaux. "Towards An Ethical Rebellion System". In: *RAD-AI @ AAMAS*. 2023.

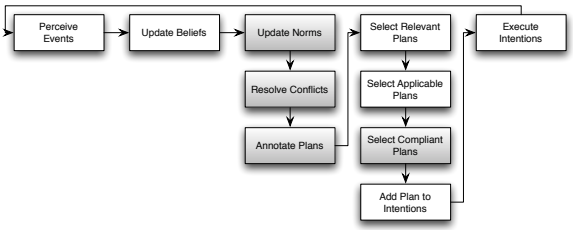
Norm Modification in BDI

Runtime Adaptation[5]

- Open environments deliver norms to agents *at runtime*
- **Obligations**: generate *new plans* to fulfil them
  - Plan creation triggered by norm activation
- **Prohibitions**: *suppress* execution of violating plans
  - Plans annotated and filtered at selection time
- This is a blunt instrument: the agent cannot disobey even when it should.

[5] Meneguzzi and Luck, "Norm-based Behaviour Modification in BDI Agents".

- Three new processes added to BDI cycle:
  1. **Update norms** (resolve conflicts)
  2. **Annotate plan library** with norm constraints
  3. **Select compliant plans** at plan selection time
- Effect of norms computed at *norm receipt*
- Compliance decision *delayed* as long as possible

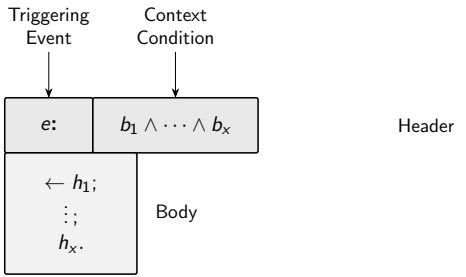


[6] Felipe Meneguzzi et al. "BDI Reasoning with Normative Considerations". In: *Engineering Applications of Artificial Intelligence* 43 (2015).

- When a specific norm activates:
  - Normative formula compared to each plan step
  - Each step annotated with appropriate constraints
- Plan header extended with a *normative header*:

$$\langle e : b_1 \wedge \cdots \wedge b_x \mid \gamma_1 \wedge \cdots \wedge \gamma_i \rangle$$

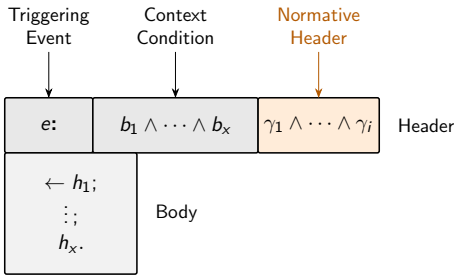
- At plan selection: check normative header
  - Compliant plan instantiation chosen where possible



- When a specific norm activates:
  - Normative formula compared to each plan step
  - Each step annotated with appropriate constraints
- Plan header extended with a *normative header*:

$$\langle e : b_1 \wedge \cdots \wedge b_x \mid \gamma_1 \wedge \cdots \wedge \gamma_i \rangle$$

- At plan selection: check normative header
  - Compliant plan instantiation chosen where possible
  - Otherwise, selective disobedience



Principled Norm Violation

When Compliance is Impossible

- What if *all* applicable plans are normatively unsatisfiable?

$$(C = \mathit{london} \wedge C \neq \mathit{london}) \rightarrow \bot$$

- Naïve response: agent fails to achieve goal
- ν-BDI response: **selective and incremental norm violation**
  - Agent identifies *which* norms must be violated
  - Chooses the plan that minimises violation
  - Norm violation is *explicit* and *reasoned*, not accidental
- This is the formal basis for *principled disobedience*[7]

Key Result

An agent can comply with norms in all cases *where possible*, and violate them only when goal achievement would otherwise be impossible.

[7] Divya Sundaresan, Gordon Briggs, and Munindar P. Singh. "Obey, Refrain, or Contravene?: Disobedience as a Basis for Responsible Autonomy". In: *AAMAS*. 2025.

Adaptive normative reasoning

Selective disobedience when planning is cheap

- **2010 assumption:** plan libraries are hand-crafted and expensive
  - ν-BDI was conservative: annotate existing plans, defer decisions
- **2025 reality:** generalised planning[8] makes new plan synthesis cheap
  - Agents can reason *from scratch* about normative constraints
- New spectrum of normative behaviour:
  1. **Comply** — synthesise a plan satisfying all norms
  2. **Minimally violate** — synthesise the plan with the fewest violations
  3. **Rebel** — explicitly ignore specific norms, with justification

Compliance becomes a *cost–benefit calculation*, not a binary constraint

[8] Pereira, Oren, and Meneguzzi, "Generalised BDI Planning".

Adaptive normative reasoning

Selective disobedience when planning is cheap

- **2010 assumption:** plan libraries are hand-crafted and expensive
  - ν-BDI was conservative: annotate existing plans, defer decisions
- **2025 reality:** generalised planning[8] makes new plan synthesis cheap(er)
  - Agents can reason *from scratch* about normative constraints
- New spectrum of normative behaviour:
  1. **Comply** — synthesise a plan satisfying all norms
  2. **Minimally violate** — synthesise the plan with the fewest violations
  3. **Rebel** — explicitly ignore specific norms, with justification

Compliance becomes a *cost–benefit calculation*, not a binary constraint
But is this *utilitarian/consequentialist* view enough?

[8] Pereira, Oren, and Meneguzzi, "Generalised BDI Planning".

Table of Contents

Reading Implicit Behaviour

Why we need to see what others don't say

- Agents (be they human, or AI/LLM) do not always reveal their goals or norms
  - **Hidden goals**: what is this agent *actually* trying to achieve?
  - **Hidden norms**: what rules govern their behaviour?
- Two mechanisms to surface implicit behaviour:
  1. **Goal recognition:** infer intent from observed actions
  2. **Norm identification:** infer governing rules from patterns of compliance and violation
- Both are prerequisites for principled rebellion:
  - You cannot disobey a norm you have not identified
  - You cannot rebel against a goal (or towards an implicit goal) you have not recognised

Reading Implicit Behaviour

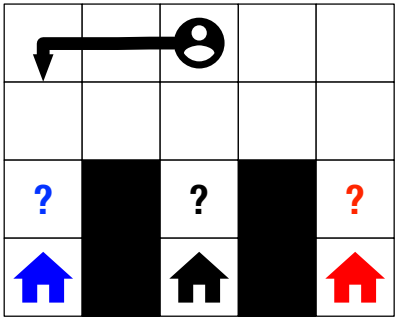
Why we need to see what others don't say

- Agents (be they human, or AI/LLM) do not always reveal their goals or norms
  - **Hidden goals**: what is this agent *actually* trying to achieve?
  - **Hidden norms**: what rules govern their behaviour (even unconsciously)?
- Two mechanisms to surface implicit behaviour:
  1. **Goal recognition:** infer intent from observed actions
  2. **Norm identification:** infer governing rules from patterns of compliance and violation
- Both are prerequisites for principled rebellion:
  - You cannot disobey a norm you have not identified
  - You cannot rebel against a goal (or towards an implicit goal) you have not recognised

Goal Recognition

Inferring intent from behaviour

- Given observations $\mathcal{O}$ and candidate goals $\mathcal{G}$: infer the *most likely* goal $g^* \in \mathcal{G}$
- **Planning-based**[9]:
  - Landmarks[10]: necessary facts in every plan for g
  - Match observed actions to landmarks (fast)
- **RL-based**[11]:
  - Q-functions encode goal proximity
  - Inference: compare Q-values (no planning, no domain model)



[9] Miquel Ramirez and Héctor Geffner. "Plan Recognition as Planning". In: *IJCAI*. 2009.

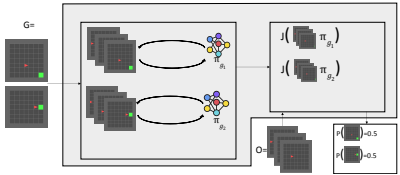
[10] Ramon Fraga Pereira, Nir Oren, and Felipe Meneguzzi. "Landmark-Based Approaches for Goal Recognition as Planning". In: *Artificial Intelligence* 285 (2020).

[11] Leonardo R. Amado, Reuth Mirsky, and Felipe Meneguzzi. "Goal Recognition as Reinforcement Learning". In: *AAMAS*. 2023.

Goal Recognition as RL

Scaling without domain models

- Q-value functions encode goal proximity: agents pursuing g increase $Q(s, a, g)$ over time
- **GRAQL**:[12] one tabular Q-function per goal
  - Inference = compare Q-values across goals
  - No domain model at test time
- **DRACO**:[13] scales to continuous domains
  - Deep Q-network per goal; partial observability, noise
  - Deployed in physical/robotic settings



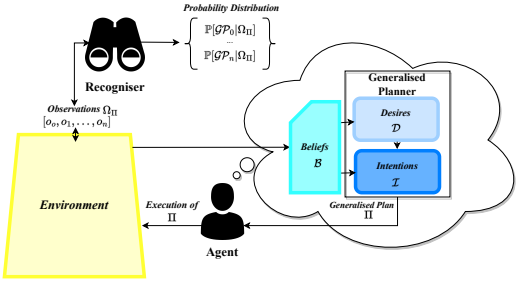
[12] Amado, Mirsky, and Meneguzzi, "Goal Recognition as Reinforcement Learning".

[13] Ben Nageris, Felipe Meneguzzi, and Reuth Mirsky. "Goal Recognition using Actor-Critic Optimization". In: *arXiv:2501.01463* (2024).

Compliance Monitoring

GR applied to norm-governed environments

- Observe another agent's actions; ask:
  - **What goal** is being pursued?
  - Does the observed plan **violate any norm**?
  - Is the violation **deliberate** or accidental?
- Supports coordination and counterplanning as by-products[a]
- Opens the adversarial question:
  - Can an agent *appear* compliant while rebelling?
  - Deceptive compliance as a research direction



[a] Michael T. Cox. "Plan Recognition for Rebel Agent Systems". In: *RAD-AI @ AAMAS*. 2025.

Learning Norms from Behaviour

- GR tells us *what* agents do; norm identification tells us *what rules* shaped it
- **Plan-recognition based**[14]: find common constraints across compliant plans
  - Obligations: actions in *every* compliant plan
  - Prohibitions: actions in *no* compliant plan
- **Bayesian**[15]: update norm hypotheses from compliant *and* violating behaviour
  - A disobedient agent's behaviour is *evidence* about the norm being violated
  - Can identify *whether* the violation is deliberate

Key insight

Deliberate rebellion requires knowing which norm you are violating. Norm identification is how an observer (or the agent itself) finds out.

[14] Nir Oren and Felipe Meneguzzi. "Norm Identification through Plan Recognition". In: *COIN@AAMAS*. 2013.

[15] Stephen Cranefield et al. "A Bayesian Approach to Norm Identification". In: *ECAI*. 2015.

Table of Contents

Principled Disobedience

A (in)formal account

1. Agent maintains a belief base and plan library (vanilla BDI)
2. Agent receives norms and annotates them into plans (ν-BDI)
3. Agent pursues goals; selects most compliant available plan
4. When *all* plans violate some norm:
   - Identify the minimal norm violation required
   - Execute with explicit record of violation
   - This is *not quite rebellion*, but rather *reasoned* non-compliance
5. Throughout all of this, agents observe behaviour and:
   - Use goal recognition to identify the pursued goal
   - Use norm identification to infer which norms emerge, and/or get violated
   - Judge whether the violation was intentional or unavoidable

Open Challenges & Ongoing Work

RAD-AI Research Agenda

- **Adversarial goal recognition**: detecting agents that *deceive*
  - Deliberately mimicking compliant behaviour
  - Deceptive planning to evade norm monitoring
- **Norm learning in multi-agent systems**
  - Norms emerge from interaction
  - Reinforcement learning of normative monitoring
- **Normative BDI + LLMs**
  - LLM-based agents: norms encoded implicitly in weights (or hidden system prompts)
  - How do we reason about norm compliance in LLM agents?
  - Value alignment as a special case of norm identification
- **Theory of Mind meets normative reasoning**
  - Recursive modelling of compliant vs. non-compliant agents

[16] Shuwa Miura. "On the Depths of Recursive Thinking in Observer-Aware Planning". In: *RAD-AI @ AAMAS*. 2024.

[17] Addison, Larue, and Molineaux, "Towards An Ethical Rebellion System".

[18] Dale Peasley, Zachary Gray, and Sandip Sen. "Disobedience for a Cause: Leveraging Implicit Objectives in User Plans by LLM Surrogates". In: *AAMAS*. 2025.

Open Challenges & Ongoing Work

RAD-AI Research Agenda

- **Adversarial goal recognition**: detecting agents that *deceive* ✓[16]
  - Deliberately mimicking compliant behaviour
  - Deceptive planning to evade norm monitoring
- **Norm learning in multi-agent systems** ✓[17]
  - Norms emerge from interaction
  - Reinforcement learning of normative monitoring
- **Normative BDI + LLMs** ∼[18]
  - LLM-based agents: norms encoded implicitly in weights (or hidden system prompts)
  - How do we reason about norm compliance in LLM agents?
  - Value alignment as a special case of norm identification
- **Theory of Mind meets normative reasoning** ✓[19]
  - Recursive modelling of compliant vs. non-compliant agents

[16] Miura, "On the Depths of Recursive Thinking in Observer-Aware Planning".

[17] Addison, Larue, and Molineaux, "Towards An Ethical Rebellion System".

[18] Peasley, Gray, and Sen, "Disobedience for a Cause: Leveraging Implicit Objectives in User Plans by LLM Surrogates".

[19] Miura, "On the Depths of Recursive Thinking in Observer-Aware Planning".

Table of Contents

Future Work

A research agenda

- Normative reasoning for LLM-based BDI agents
- Adversarial goal recognition and deceptive compliance
- Norm identification from multi-agent observation

Future Work

A research agenda

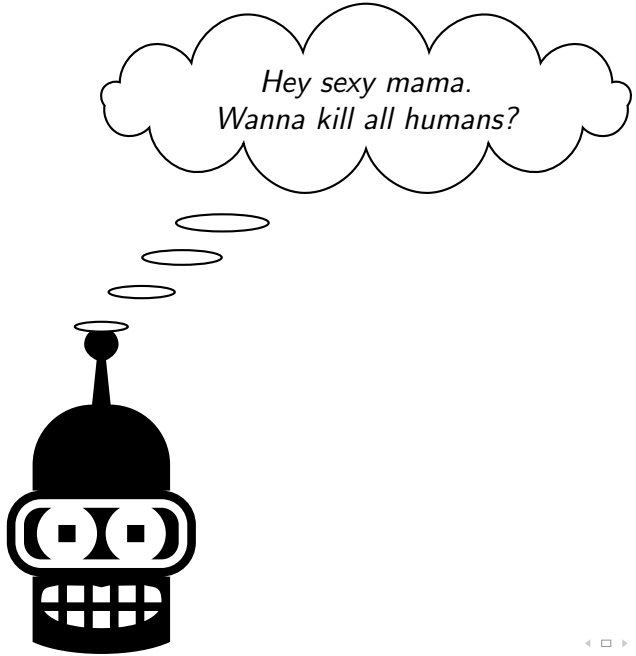
- Normative reasoning for LLM-based BDI agents
- Adversarial goal recognition and deceptive compliance
- Norm identification from multi-agent observation

The game is out there.
It's either play, or get played.

Acknowledgements

Not my genius alone

- **BDI / Generalised Planning**: Ramon Fraga Pereira, Nir Oren
- ν-**BDI**: Odinaldo Rodrigues, Michael Luck, Wamberto Vasconcelos, Nir Oren
- **Goal Recognition**: Ramon Fraga Pereira,Leonardo Amado, Ben Nageris, Mor Vered, Miquel Ramirez, Reuth Mirsky, and **many others**
- **Norm Identification**: Nir Oren, Stephen Cranefield
- Funding: EPSRC, CNPq, FAPERGS, Google


Hey sexy mama.
Wanna kill all humans?

Thank you!

felipe.meneguzzi@abdn.ac.uk

UNIVERSITY OF
ABERDEEN