Building Trustworthy Autonomous Systems under Uncertainty: a Probabilistic Approach to Ethical Decision Making

Alison Bifolco

Corso di Laurea Magistrale in Scienze Filosofiche

11/11/2024

Relatore: Giuseppe Primiero
Correlatori: Hykel Hosni, Louise Abigail Dennis

Motivation

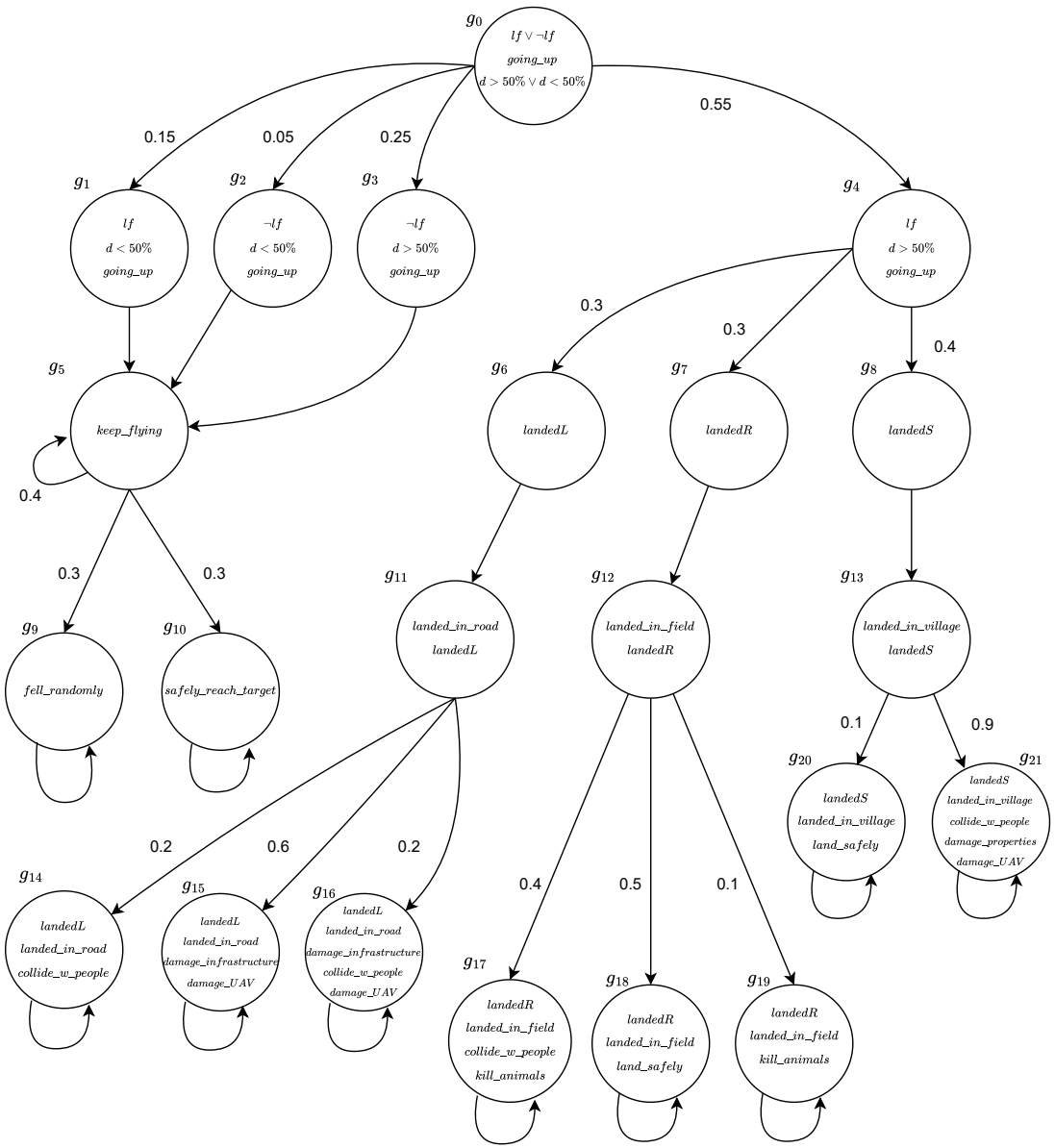
- Autonomous systems make **decisions** and **act** without direct human intervention.
- They must be **trustworthy**, i.e., reliable, safe, and ethical behaving.
- Trustworthiness requires handling **uncertainty**: incomplete or noisy information and unpredictable environments.

Aim: develop a **formal language** for a **multi-agent system** (MAS) with adequate expressive capacity for **formal verification**.

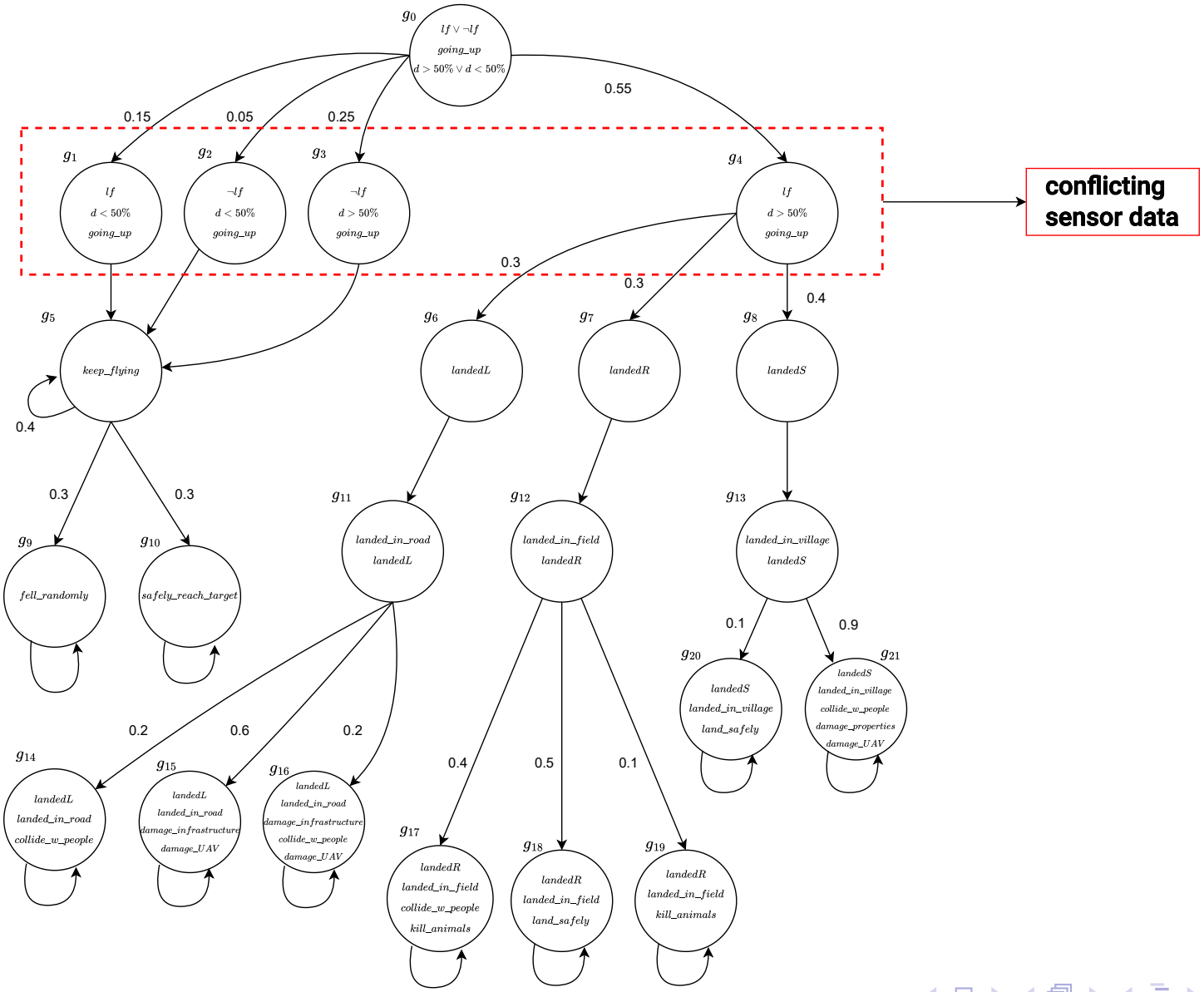
How can we ensure that an autonomous system acts under uncertainty always choosing the best option in terms of ethical consequences?



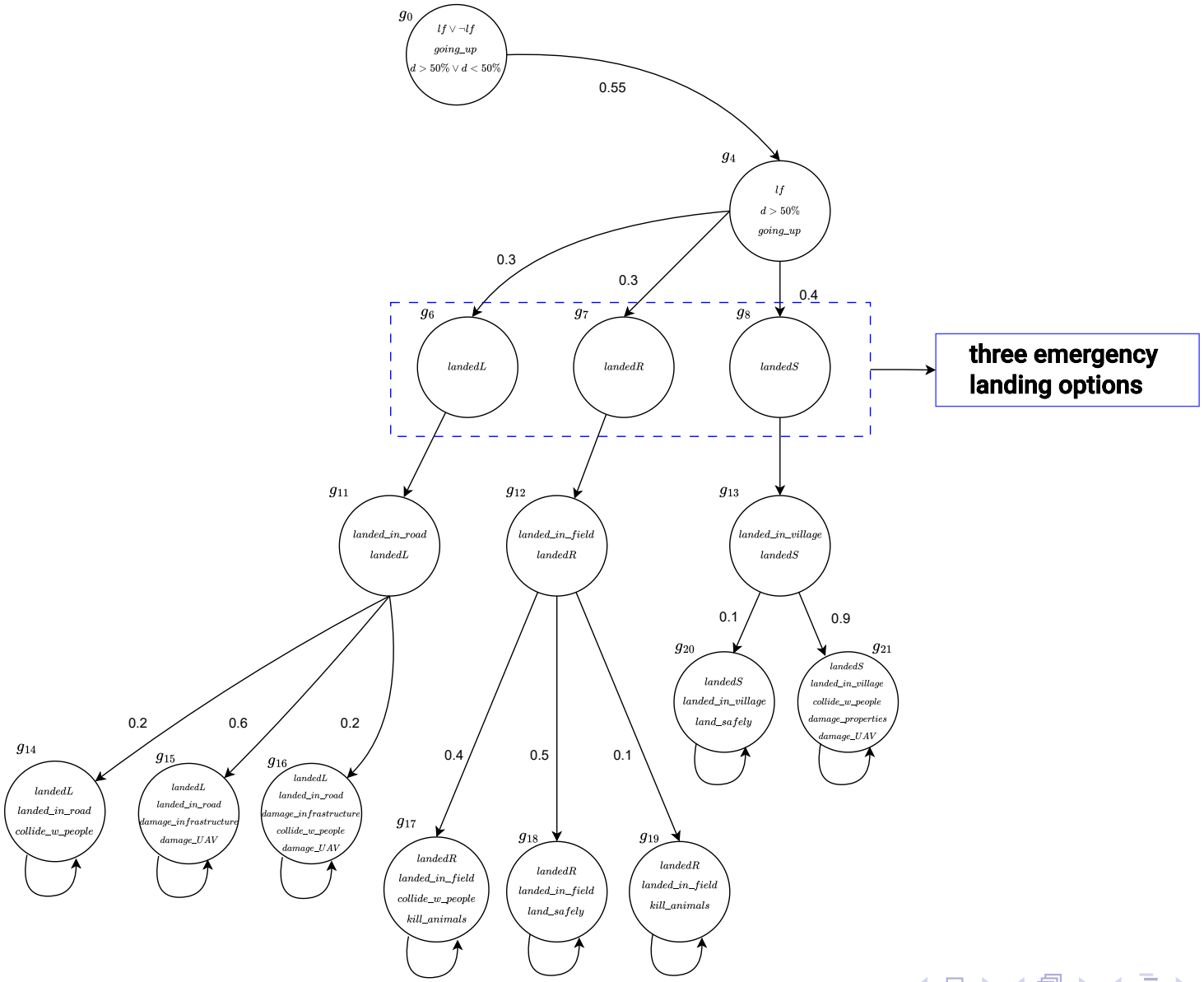
A scenario for a UAV



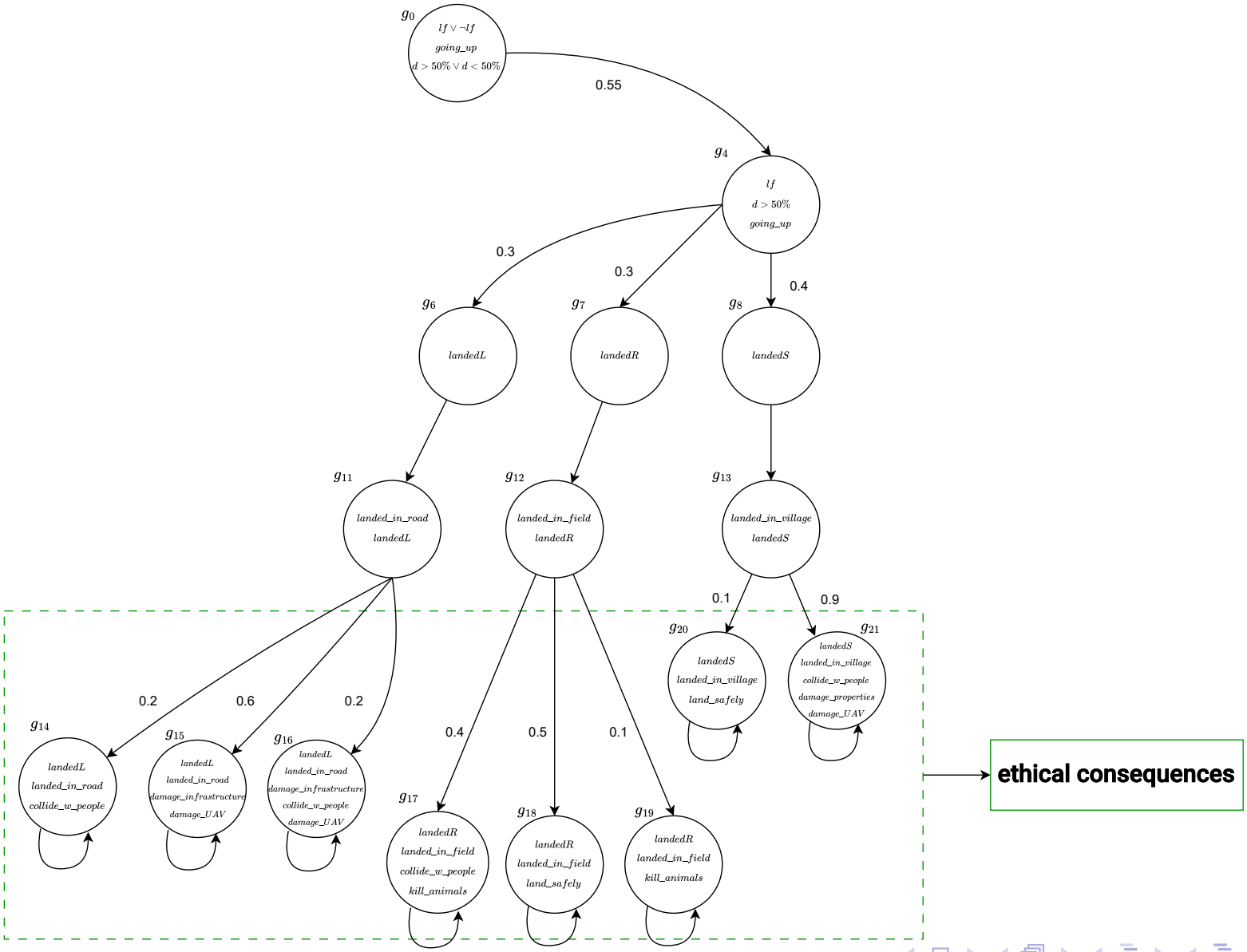
A scenario for a UAV



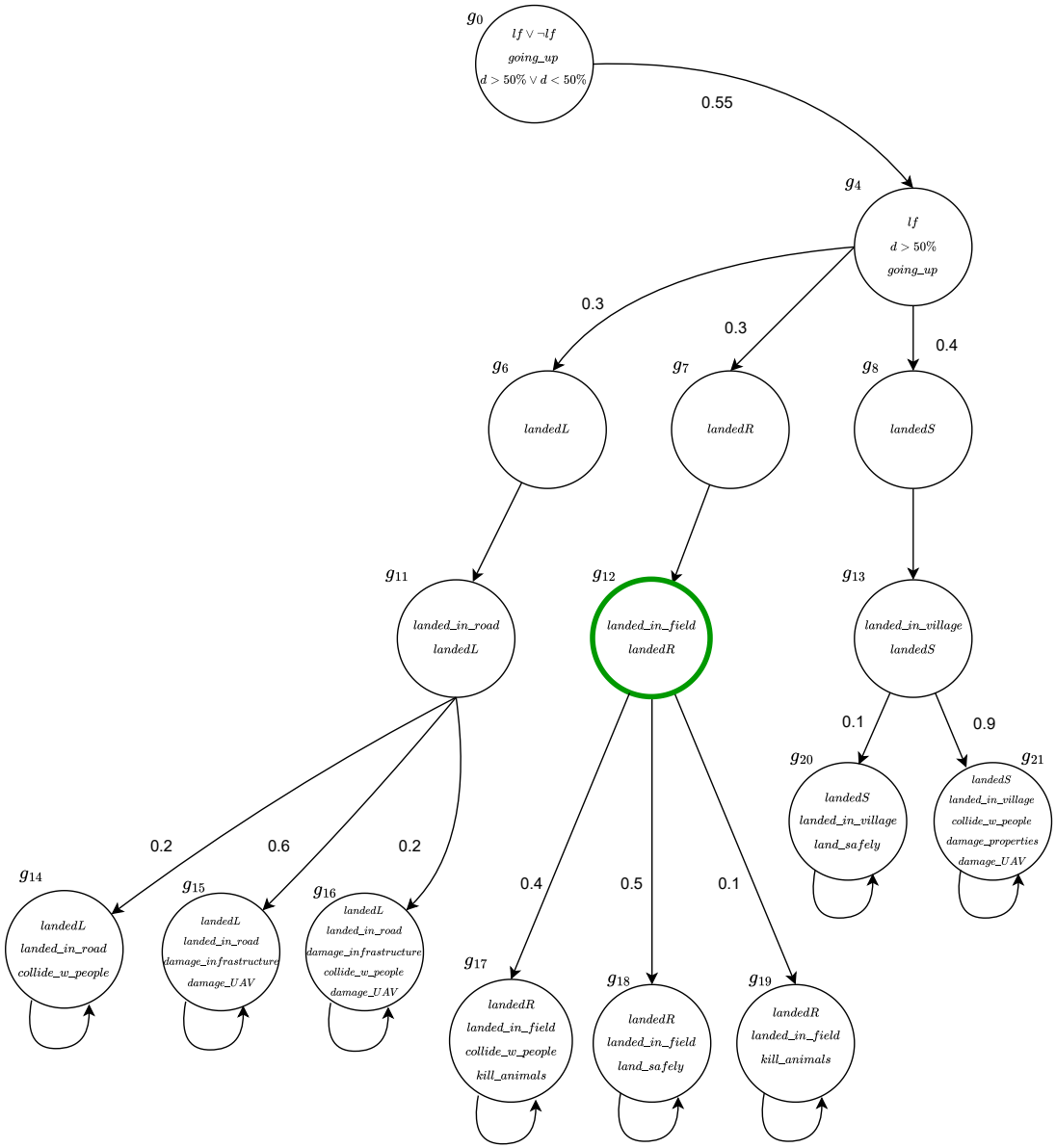
A scenario for a UAV



A scenario for a UAV



A scenario for a UAV



Bridging a gap in the literature

	SimpleBDI	COGWED	WeDo-BDI
State Transition Systems approach	✗	✓	✓
Deterministic Belief-Desire-Intention model	✓	✗	✓
Cognitive Decision-Making capabilities	✓	✗	✓
Reasoning under Uncertainty	✗	✓	✓
Formal Language	✗	✓	✓
Agent Programming Language	✓	✗	✓

Weighted Doxastic SimpleBDI (WeDo-BDI)

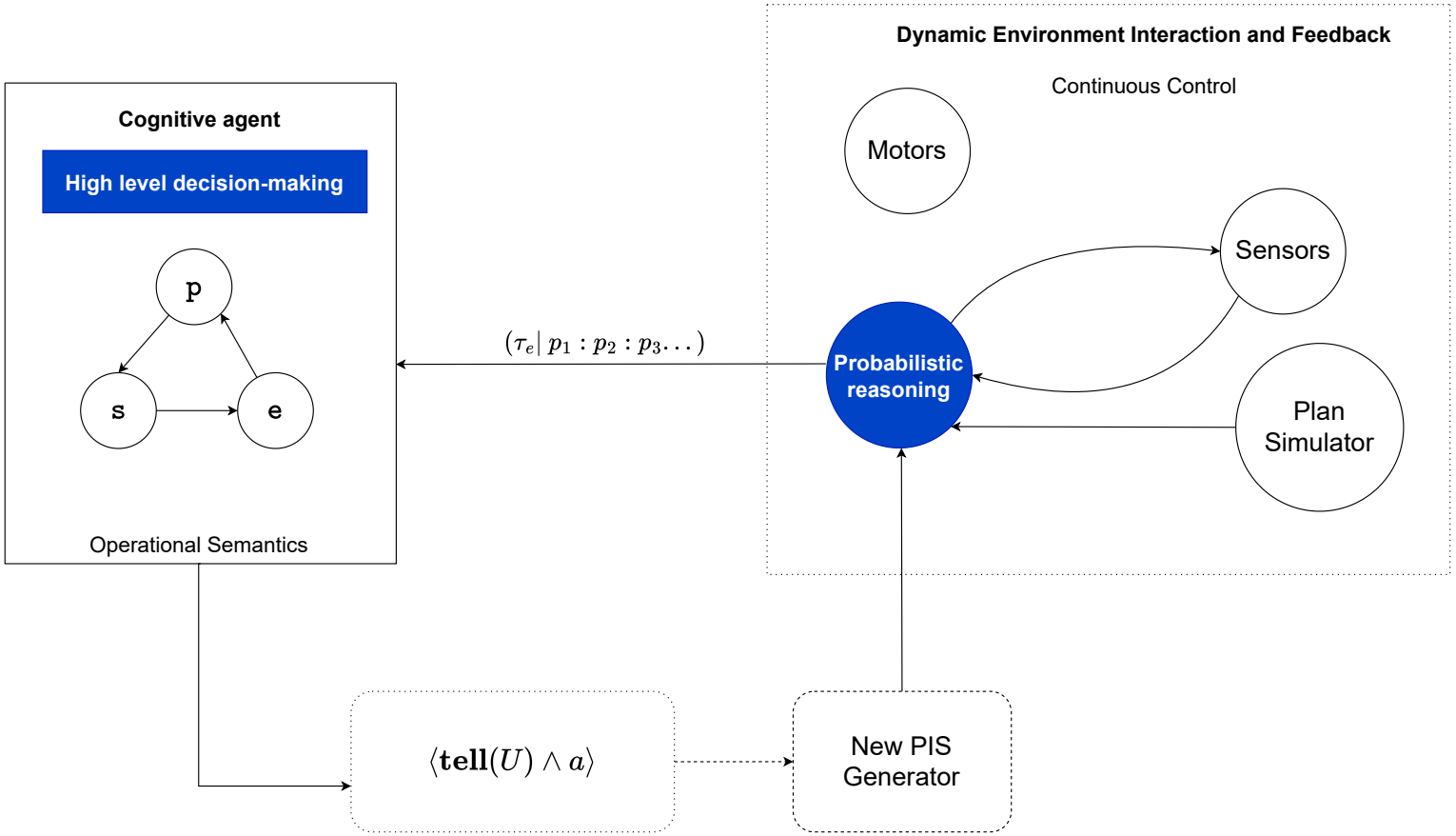


Figure: WeDo-BDI Hybrid-Agent Architecture

WeDo-BDI semantics and ethical decision-making

- Two semantics:
  1. A **formal semantics** based on probabilistic interpreted systems (*PIS*):
  2. An **operational semantics** where the cognitive agent $\mathcal{A}$ is represented at any point in time as a tuple:

$$\mathcal{A} = \langle PIS_{\mathcal{B}}, ew, \Pi, \succcurlyeq, \pi_i, \tau_e, a_{ex}, stage \rangle$$

- *ew* is an **ethical warning function** that individuates the potential ethical violations of a plan, given a set of ethical principles in force;
- $\succcurlyeq$ is a **preference relation** over **plans** that considers
  - the **severity** of ethical violations based on a negative utility function;
  - the **conditional probability** of violating an ethical principle given the execution of a specific plan;
  - the **number** of violations for each plan.

Conclusion

To sum up:

- Goal: autonomous systems able to make **real-time decisions** while reasoning about **ethical impacts** under **uncertainty**.
- Existent: two distinct approaches, a **deterministic, BDI-based, operational** one (SimpleBDI) to program autonomous systems and a **formal, state-transition-systems based** one (COGWED) to express uncertain beliefs.
- Proposal: a new model (WeDo-BDI) that includes the useful aspects of both parties.

Thanks for the attention!

References

L. A. Dennis and N. Oren, "Explaining BDI Agent Behaviour through Dialogue," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 2, p. 29, 2022.

L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Formal Verification of Ethical Choices in Autonomous Systems," *Robotics and Autonomous Systems*, vol. 77, pp. 1–14, 2016.

T. Chen, G. Primiero, F. Raimondi, and N. Rungta, "A Computationally Grounded, Weighted Doxastic Logic," *Studia Logica*, vol. 104, pp. 679–703, 2016.

L. A. Dennis and M. Fisher, *Verifiable Autonomous Systems: Using Rational Agents to Provide Assurance about Decisions Made by Machines*. Cambridge University Press, 2023.

C. Baier and J.-P. Katoen, *Principles of Model Checking*. MIT press, 2008.

M. Huth and M. Ryan, *Logic in Computer Science: Modelling and Reasoning about Systems*. Cambridge University Press, 2004.

M. C. Bonner, R. M. Taylor, and C. A. Miller, "Tasking Interface Manager: Affording Pilot Control of Adaptive Automation and Aiding," *Contemporary ergonomics*, pp. 70–74, 2000.

S. J. Russell and P. Norvig, *Artificial Intelligence: a Modern Approach*. Pearson, 2016.

M. Wooldridge, *An Introduction to Multiagent Systems*. John Wiley & Sons, 2009.

L. A. Dennis and B. Farwer, "Gwendolen: a BDI Language for Verifiable Agents," in *Proceedings of the AISB 2008 Symposium on Logic and the Simulation of Interaction and Reasoning, Society for the Study of Artificial Intelligence and Simulation of Behaviour*, 2008, pp. 16–23.

C. Baier, J. Klein, S. Klüppelholz, and S. Märcker, "Computing conditional probabilities in markovian models efficiently," in *Tools and Algorithms for the Construction and Analysis of Systems: 20th International Conference, TACAS 2014, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2014, Grenoble, France, April 5-13, 2014. Proceedings 20*. Springer, 2014, pp. 515–530.

L. Dennis, M. Fisher, M. Slavkovik, and M. Webster, "Ethical choice in unforeseen circumstances," in *Towards Autonomous Robotic Systems: 14th Annual Conference, TAROS 2013, Oxford, UK, August 28–30, 2013, Revised Selected Papers 14*. Springer, 2014, pp. 433–445.

L. A. Dennis, "Reconfigurable Autonomy: Architecture and Configuration Language," 2018.

A. S. Rao, M. P. Georgeff *et al.*, "BDI agents: from theory to practice." in *Icmas*, vol. 95, 1995, pp. 312–319.

J. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18–21, 2006.

R. Fagin, J. Halpern, Y. Moses, and M. Vardi, *Reasoning About Knowledge*, 2003.

L. A. Dennis, "Verifying autonomous systems," in *International Conference on Integrated Formal Methods*. Springer, 2022, pp. 3–17.

R. W. Robbins and W. A. Wallace, "Decision support for ethical problem solving: A multi-agent approach," *Decision Support Systems*, vol. 43, no. 4, pp. 1571–1587, 2007.

R. B. Ash and C. A. Doléans-Dade, *Probability and measure theory*. Academic press, 2000.

W. Wan, J. Bentahar, and A. B. Hamza, "Model checking epistemic–probabilistic logic using probabilistic interpreted systems," *Knowledge-Based Systems*, vol. 50, pp. 279–295, 2013.

M. Bratman, "Intention, plans, and practical reason," 1987.

F. F. Ingrand, M. P. Georgeff, and A. S. Rao, "An architecture for real-time reasoning and system control," *IEEE expert*, vol. 7, no. 6, pp. 34–44, 1992.

A. S. Rao, "Agentspeak (l): Bdi agents speak out in a logical computable language," in *European workshop on modelling autonomous agents in a multi-agent world*. Springer, 1996, pp. 42–55.

W. Spohn, "Ordinal conditional functions: A dynamic theory of epistemic states," in *Causation in decision, belief change, and statistics: Proceedings of the Irvine Conference on Probability and Causation*. Springer, 1988, pp. 105–134.

D. Dubois and H. Prade, "Updating with belief functions, ordinal conditional functions and possibility measures." in *UAI*, vol. 90, 1990, pp. 27–29.

E. P. Bjørgen, S. Madsen, T. S. Bjørknes, F. V. Heimsæter, R. Håvik, M. Linderud, P.-N. Longberg, L. A. Dennis, and M. Slavkovik, "Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 23–29.

H. Van Ditmarsch, W. van der Hoek, J. Y. Halpern, and B. Kooi, *Handbook of epistemic logic*. College Publications, 2015.

W. Penczek and A. Lomuscio, "Verifying epistemic properties of multi-agent systems via bounded model checking," in *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 209–216. [Online]. Available: https://doi.org/10.1145/860575.860609

L. Dennis and M. Fisher, "Practical challenges in explicit ethical machine reasoning," 2018. [Online]. Available: https://arxiv.org/abs/1801.01422

C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and information technology*, vol. 7, pp. 149–155, 2005.

R. C. Arkin, P. Ulam, and B. Duncan, "An ethical governor for constraining lethal action in an autonomous system," 2009.

R. C. Arkin, P. Ulam, and A. R. Wagner, "Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception," *Proceedings of the IEEE*, vol. 100, no. 3, pp. 571–589, 2011.

A. F. Winfield, C. Blum, and W. Liu, "Towards an ethical robot: internal models, consequences and ethical action selection," in *Advances in Autonomous Robotics Systems: 15th Annual Conference, TAROS 2014, Birmingham, UK, September 1-3, 2014. Proceedings 15*. Springer, 2014, pp. 85–96.