Interpretability, Trust, and Morality in Autonomy

David Danks

Philosophy & Psychology

Carnegie Mellon

Carnegie Mellon University

- When can someone ethically use/deploy an autonomous system?

What do I mean by ‘autonomy’?

- Characterize ‘autonomy’ in terms of *capabilities* (= context-specific abilities)
  - Planning (routes, action sequences, …)
  - Learning (environmental statistics, adaptation, …)
  - Deciding (action selection, classification, …)
  - …
- Richer & more useful than ‘levels’
  - Though obviously also more complicated…

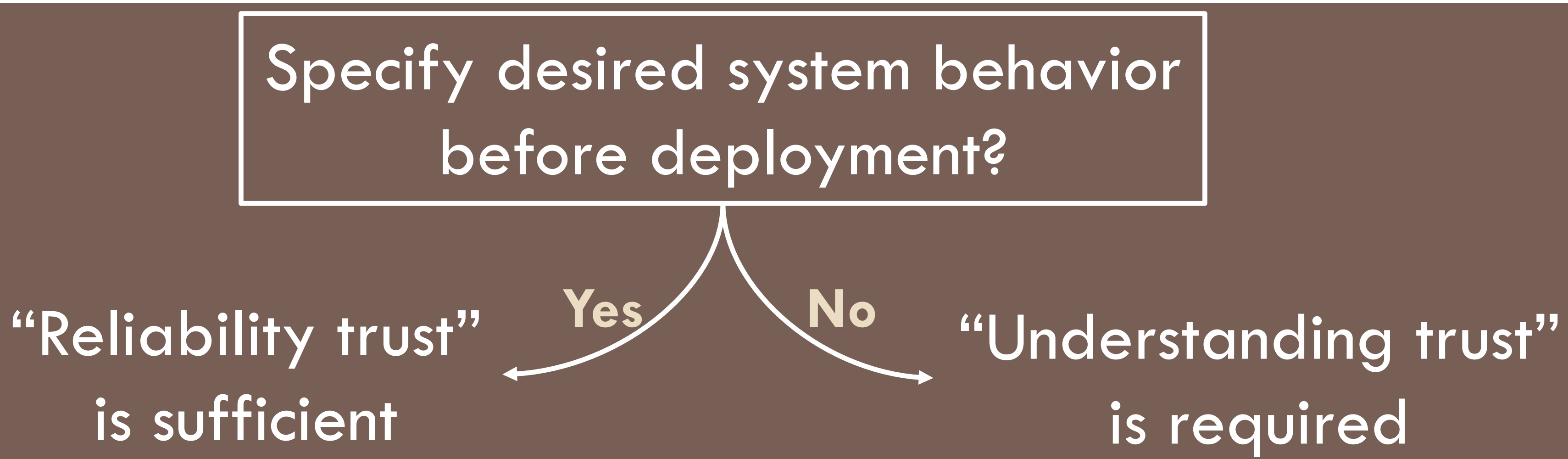
Ethical deployment

- *Necessary condition for ethical deployment*: Reasonable belief that the "system" will behave (approximately) as the user intends
  - Agnostic about whether "system" is human or artificial
  - User's intentions can be quite high-level ("drive safely")
- *Alternate formulation*: User must trust the "system"
  - In *relevant* respects, not necessarily all of them

Varieties of (psychological) trust

1. **Reliability/Predictability** of Trustee
   - “Behavior”-focused
   - Knowing *what* the trustee will do
   - Coordination & prediction in known circumstances
2. **Understanding** the Trustee
   - Belief/value-focused
   - Knowing *why* the trustee will do
   - Coordination & prediction in novel circumstances

Trust & ethical deployment



Specify desired system behavior before deployment?

Yes

“Reliability trust” is sufficient

No

“Understanding trust” is required

- ⇒ If system will use autonomous capabilities, then deployer must have “understanding trust”

Trust & interpretability

- *Interpretability* is key for “understanding trust”
- ⇒ Necessary condition for ethical deployment is: “System behavior for Goal is interpretable by User”
  - *Note*: Interpretability is *not* a property of System alone

Trust & interpretability

- Routes to interpretability & "understanding trust"
  1. Explicit requirement that System plan/learn/decide similarly to humans
  2. Have deployment decisions made in collaboration with (informed) developers
  3. Extended user experience in many different contexts

What about system morality?

- *Observation*: Humans typically interpret others' unethical behavior using "internal" features
  - *Conjecture*: Non-developers will usually interpret unethical system behavior as due to an unethical nature

- *Conjecture*: If the developer cares about having an ethical system, then any u-trustworthy system will (mostly) act ethically

Conclusions

- Ethical deployment requires trust (in system)
- If a system employs autonomous capabilities, then understanding-trust is required
- Interpretability is necessary for understanding-trust
  - Reminder: ‘interpretability’ is a three-argument relation
- ⇒ Interpretability is nec. for ethical deployment
  - Though many routes to this type of interpretability
- *Conjecture*: Alternate route to try to achieve ethical system behavior?

Thanks!