FROM VALUES TO CONSTRAINTS TO ASSURANCE

DAVID DANKS

DATA SCIENCE // PHILOSOPHY
UNIVERSITY OF CALIFORNIA, SAN DIEGO

ASSURANCE DEPENDS ON VALUES

- Autonomous systems are not fully predictable
  - Assured autonomy cannot be (purely) about reliability

- Instead, assurance is about the system supporting our values

- But even if we know the relevant values, how can we use them to achieve *the right kind of* assurance?

FROM VALUES TO CONSTRAINTS

- Values imply (but are not equivalent to) behavioral constraints
  - “Safe driving” ⇒ “do not speed”, “do maintain awareness”, etc.
  - “Fair hiring” ⇒ “do not consider race”, “do determine skills”, etc.

- Constraints can be:
  - Inexact
  - Context-sensitive
  - Use-appropriate
  - Expressible in different languages
  - …

Describe required & forbidden behavior
Leave other aspects unresolved

FROM CONSTRAINTS TO ASSURANCE

- Given a system-model, we can then prove / simulate:
  - Any constraints that are always violated
  - Contexts that might produce a constraint violation
  - Potential incompatibilities between constraints
  - Dynamics that potentially threaten to violate a constraint

- If the constraints accurately "track" the values, then we have a path to assurance