IFIP 10.4 11 May 2022

Models and their Validation and their Role in Perception And in Safe Autonomous Vehicles

John Rushby

Computer Science Laboratory
SRI International
Menlo Park, CA

Models

- A model is a simplified description of something
    - To be used in explaining or predicting behavior of the actual thing
- Their use is probably as old as our species
- Maybe mythical/supernatural at first: thunder is caused by the gods
    - Prediction errors attributed to faulty interpretation rather than poor model (e.g., Oracle of Delphi)
- But some were geometric and quasi-mathematical: astronomy, Archimedes principle
    - Predict seasons, eclipses, conjunctions, etc.
    - Preference for explanatory models (Aristotle)
    - But also smaller prediction errors (Ptolemy)

Validation (Aristotle)

- Want an argument from observation to explanation (i.e., model)
- Notion of a demonstrative argument (sets a high bar)
  - A valid argument that does not merely show the conclusion is true
  - But also why it is true (i.e., explanation)
  - All the propositions in the argument must be necessary, general and eternal truths
  - And must bottom out in fundamental certain truths (grasped by senses)
- Modern criticism: validation proceeds anti-causally
  - The world causes our sensations
  - But demonstration reasons from sensations to (our model of) the world

    So it goes backwards, from the caused to the cause
- 2,000 years go by...

Validation (Galileo)

- The dawn of (European) science
  - Saw things never seen before (mountains on moon, sunspots, moons of Jupiter)
  - Built mathematical models (velocity of falling body)
- Validation by "Method of Regress," due to Zabarella (Padua)
  - From observation, try to discern cause
  - Then demonstrate that the cause leads to the observation
  - Not circular: intermediate step of considering and testing the cause allows us to understand the observation differently than before
- Modern criticism: an improvement on Aristotle, but still partially anti-causal
- Galileo also used falsification: e.g., phases of Venus falsify geocentric cosmology
- However, al-Haytham (circa 1,000) was closer to modern scientific method

Validation (Scientific Method)

- Massive progress in science and engineering from Galileo to present
- But little change in validation: preference for small prediction errors, simplicity and explanation; falsification acknowledged
  - This is challenged by models that predict but do not explain
  - "Those not shocked by quantum theory cannot possibly have understood it" (Bohr)
- Coherent treatments of validation are quite recent (Peirce, Vienna School)
  - Explicit formulation and experimental testing of hypotheses (i.e., models, theories)
  - Testing means looking for and evaluating prediction errors
  - Falsification can be seen as extreme prediction errors
  - Science is identified with (defined by) models that are potentially falsifiable (Popper)
- But that's not how science is done: paradigms withstand prediction errors until... (Kuhn)
  - Science advances one funeral at a time (Planck)
- Scientists appreciate Popper, philosophers appreciate Kuhn

Modern Validation by Scientific Method

- Appreciation: validation proceeds causally
  - The world causes our observations
  - We use our model (of the world) to predict observations
  - Then contemplate prediction error (or falsification if extreme)
- Confirmation Theory
  - Consider conditional probability of an observation given one model vs. another
  - Gives rise to Confirmation Measures (Carnap, Hempel)
  - Applied to assessment of "weight" of evidence (Good & Turing)
  - And then to software assurance

    Google "rushby biblio" then it's the latest (2022) tech report
- Aside: humans evolved to "weigh" evidence and it's likely we use confirmation measures rather than basic probabilities (cf. conjunction "fallacy" of Kahneman & Tversky)

Further Aside: Models in Science and Engineering

- In his excellent book "Plato and the Nerd" Ed Lee observes that models are used in diametrically opposite ways in science and engineering
- In science, models are descriptive and predictive: they tell how the world is
- In engineering, models are prescriptive: they tell how the world should be
- Problems if you confuse the two
- Further-Further Aside: in assurance
  - You want a descriptive model of the world with the system in it, check it is safe etc.
  - The designers had a prescriptive model of the system
  - If you can show the actual system matches that model (i.e., verification)
  - Then can use prescriptive system model plus descriptive world model to check safety

Model-Based Control

- Conant & Ashby: "Every Good Regulator of a System Must be a Model of that System"
- In classical control theory,

  A model is used in design but is not explicitly present in the operational system
- But as the world changes, so ought the model, then needs to be present in operational system
  - e.g., pressure and temperature of atmosphere change as plane climbs
  - So allow controller parameters to be adjusted during operation (adaptive control)
  - Problem of validating the changing model at runtime
  - cf. fatal crash of X-15: problem not in adaptive control considered alone, but in presence of other failures—80% loss of effectiveness in one elevon
  - FAA: Use multiple validated models and move between them (gain scheduling)
- May have uncertainty regarding the plant or environment
  - Make more of the model explicit in the system
    - ⋆ And subject to modification or construction during operation
  - But again, how do we validate the current model?
    - ⋆ Maybe apply modern FDIR: adjust model until predictions match observations

(Human) Perception

- Perception is about building a representation of the world (i.e., a model)
- That is useful for prediction (i.e., for planning actions)
- Let's consider vision
- Early theories had "rays" coming out of the eyes and acting like a remote sense of touch
- Now, we at least have the optics right (al-Haytham again)
- But untutored view is that perception is built bottom up
  - From distorted fuzzy pixels
  - Through many levels of feature and object detection
  - To the perceived image (i.e., model) of the world
    - ⋆ In humans it is conscious (so that we can report it), but doesn't have to be (cf. "blindsight") and probably is not in animals
- Modern view: first steps OK, last one (bottom up construction of final image) not so
  - Not enough information in instantaneous "snapshot": need to integrate as a model
  - Too much ambiguity (many worlds could cause same image)

Modern Theories of Perception

- Helmholtz (1867) "Handbuch der Physiologischen Optik III"
- Gregory (1980) "Perceptions As Hypotheses" explicit comparison to scientific theories
- Clark, Hohwy, Friston: predictive coding, predictive error minimization, free energy
- Dominant theory in psychology, much of cogsci
  - We have a hierarchy of models, descending from "upper levels" into the senses
  - At each level, model predicts what lower one will perceive next
  - Prediction error is used to adjust the model at each level
    - ⋆ Big error ("surprise," or falsification) causes major reevaluation (System 2)
    - ⋆ Small errors lead to model refinement
      - ⋄ Conceptually, a Bayesian update, mechanized as iterative optimization
- Evidence:
  - Optical illusions (cf. waking up in an unfamiliar room)
  - Sight restored in adulthood to those born blind
  - More neural pathways go "down" (predictions) than "up" (prediction errors)

Validation of Perception

- Evolution must ensure that our perceptions are adequately comprehensive and correct
- Comprehensive is a function of an animal's life style
  - A frog surrounded by (edible) dead flies will starve: sees only moving ones
  - We cannot sense magnetic fields, ultra-violet
- Correctness is universal: does the perceived model enable useful predictions?
- Validation through real-life decisions may be too infrequent and too late
- So should constantly validate correctness
  - i.e., make predictions and check them
  - Makes no sense to separate validation of model from its construction
  - Hence evolution of predictive processing (just so story)
- Higher-level cognitive functions work in a similar way: mental models (Craik)

Application to Autonomous Systems/Vehicles

- System builds model of its environment
  - And uses that to plan and execute actions to achieve some goal
  - Can check safety of proposed actions, given the model
  - Aside: can also manage classical fault-tolerance of these mechanisms
- But the model is critical: has to be reasonably accurate
- System has perception based on cameras, lidar, radar, ultrasound etc.
  - Feature and object detection largely based on machine learning
    - ⋆ Known to be flawed and unreliable
  - And model is then typically built by "fusion" on these
- Criticism
  - This is anti-causal (bottom up), has all the problems associated with that
  - Concern focuses on machine learning and lower level perception
  - But the model accumulates and integrates perception, and should be the focus
  - Yet there is no validation of the model

Model Validation

- Does the model enable accurate predictions?
- Validation through real-life decisions (i.e., driving) may be too infrequent and too late
- Testing cannot get there
- So correctness of the model should be constantly validated
  - i.e., make predictions and check them
  - Makes no sense to separate validation of model from its construction
  - So change to a predictive processing architecture
- Predictive processing
  - Use model to predict output from some stage of the perception pipeline
  - Compare prediction to actual output
  - Use prediction error to adjust the model
    - ⋆ Big error ("surprise," or falsification) causes major reevaluation ("System 2 intervention")
    - ⋆ Small errors lead to model refinement
      - ⋄ Conceptually, a Bayesian update, mechanized as iterative optimization

Predictive Processing in Autonomous Vehicles

- The model and the output of perception pipeline may both use same representation
    - Typically detected objects list (what each object is, size, velocity, intent, etc.)
    - Plus occupancy grid (bird's eye view of road and object layout)
- In which case, prediction is just time-advanced model
- So prediction error is simply difference between
    - Current output of perception pipeline and time-advanced model
- And Bayesian update to model is a form of sensor fusion similar to Kalman filter
- Note that persistence of model masks intermittent perception faults.
- How can it fail? Systematic defects:
    - Perception system is blind to red cars (actually some unnamed aspect of reality)
    - Model contains no red cars, predicts no red cars
    - No prediction errors. . . until you collide with a red car
    - But if another car pulls in front of red car, it will vanish from view (occluded by red car)
    - That will trigger surprise, so not all is lost
    - But can we do better?

Diversity as Protection for Systematic Defects

- Maintain two perception pipelines, same sensors, different ML architectures, training data
  - Either feed both pipelines into same model
  - Or use one as main model, other for checking safe actions
    periodically swap them to avoid divergence
- Or add specialized perception pipelines
  - Looking for specifically for hazards such as imminent collisions
  - And feed those into the model
  - Will trigger surprise when hazard detected that main pipeline missed

Conclusion

- It's been a long journey, from Aristotle to the present
- But we end up in fairly familiar territory
- Use a perception pipeline to build a model
- Novelty is how perception updates the model: model is given more weight than usual
  - Equivalent to using prediction error to construct and validate model
  - Conventionally, output of perception pipeline is the model
- Can use redundancy and diversity of perception to tolerate systematic faults
- And I think the approach can support a plausible case for assurance
- Much of this is in SafeComp 2020 paper with colleagues Susmit Jha and Shankar
  "Model-Centered Assurance For Autonomous Systems" (on my web site)