Towards Safe AI for Learning-enabled Robots

24th November 2021

Subramanian Ramamoorthy

School of Informatics, The University of Edinburgh

Edinburgh Centre for Robotics

Alan Turing Institute

http://rad.inf.ed.ac.uk/

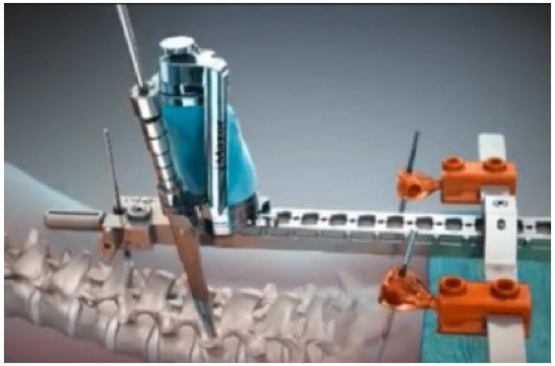
EDINBURGH CENTRE FOR ROBOTICS

THE UNIVERSITY of EDINBURGH informatics

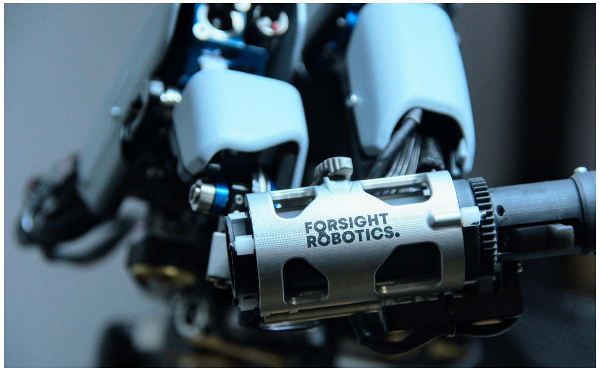
The Alan Turing Institute

A Long-term Vision:
Autonomy in Medical Robots

- The use of robots in medicine in becoming increasingly more common – ranging from “nurse-bots” to surgery



[Mazor Robotics/Medtronic]



[Forsight Robotics]

- Outside of narrow domains within diagnostic imaging, relatively little use of *AI for Autonomy* in medical applications

Autonomy in the Operating Theatre

Level 1: Assistance	Level 2: Partial Automation	Level 3: Conditionally autonomous	Level 4: Highly automated	Level 5: Full autonomy
Basic control systems	Instrumented tools, still largely in remote control paradigm, e.g., Intuitive Surgical	?	?	Undesirable, perhaps

This is the status quo for "robot surgery" currently

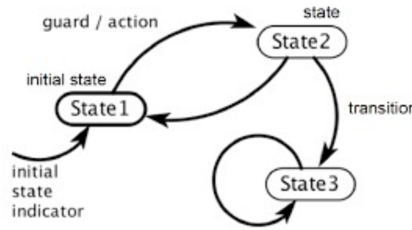
autonomy for assistance; skill prostheses

Many reasons why:

- [NHS] Shortage of (junior) staff, leading to missed targets
- [Global Health] Can a moderately skilled person do expert level work?
- Improved outcomes in terms of accuracy and time; lower lifecycle costs

Turing Project: A Technology Stack

Collab.: Dr P Brennan (NHS Lothian, UoE Clinical Brain Sciences)
Informatics PDRAs: Dr Michael Burke, Dr Craig Innes



Specifications :
learnt from data +
extracted from codes of practice
(Amenable to reasoning about safety)

$$
\begin{array}{lcl}
(\mathbf{x},t) \models \mu & \Leftrightarrow & f(x_1[t],\ldots,x_n[t]) > 0 \\
(\mathbf{x},t) \models \varphi \wedge \psi & \Leftrightarrow & (x,t) \models \varphi \wedge (x,t) \models \psi \\
(\mathbf{x},t) \models \neg\varphi & \Leftrightarrow & \neg((x,t) \models \varphi) \\
(\mathbf{x},t) \models \varphi \,\mathcal{U}_{[a,b]}\, \psi & \Leftrightarrow & \exists t' \in [t+a, t+b] \text{ such that } (x,t') \models \psi \wedge \\
 & & \quad \forall t'' \in [t,t'],\ (x,t'') \models \varphi\}
\end{array}
$$

Guaranteed
control synthesis
at Levels 3-4

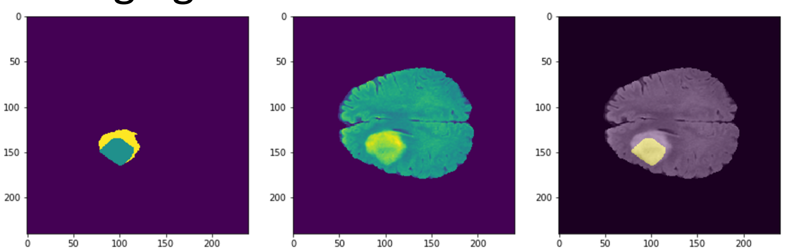
Programming
by discussion &
model induction

A Simple Control Problem

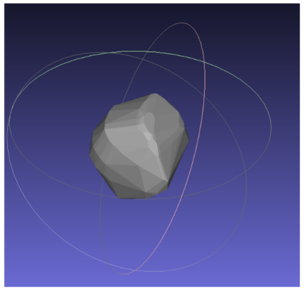
Thought Experiment:
How to Excise a Tumour?

A hypothetical pipeline:

Imaging



3D scene understanding

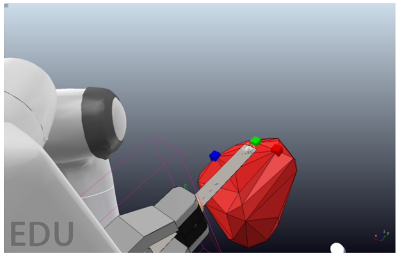


Meshing

Solve travelling salesman problem visiting faces

Path planning

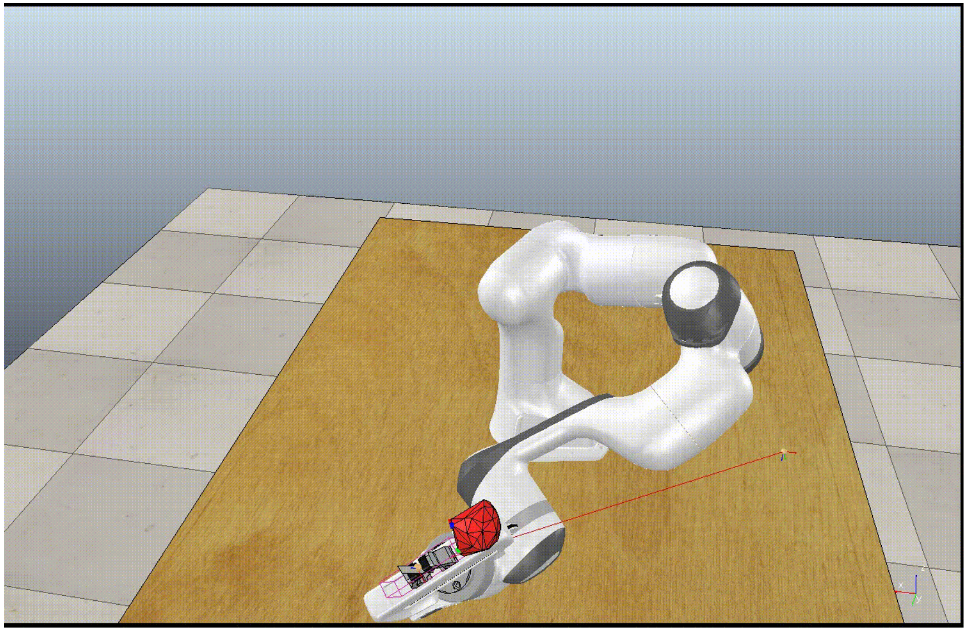
Tracking Control



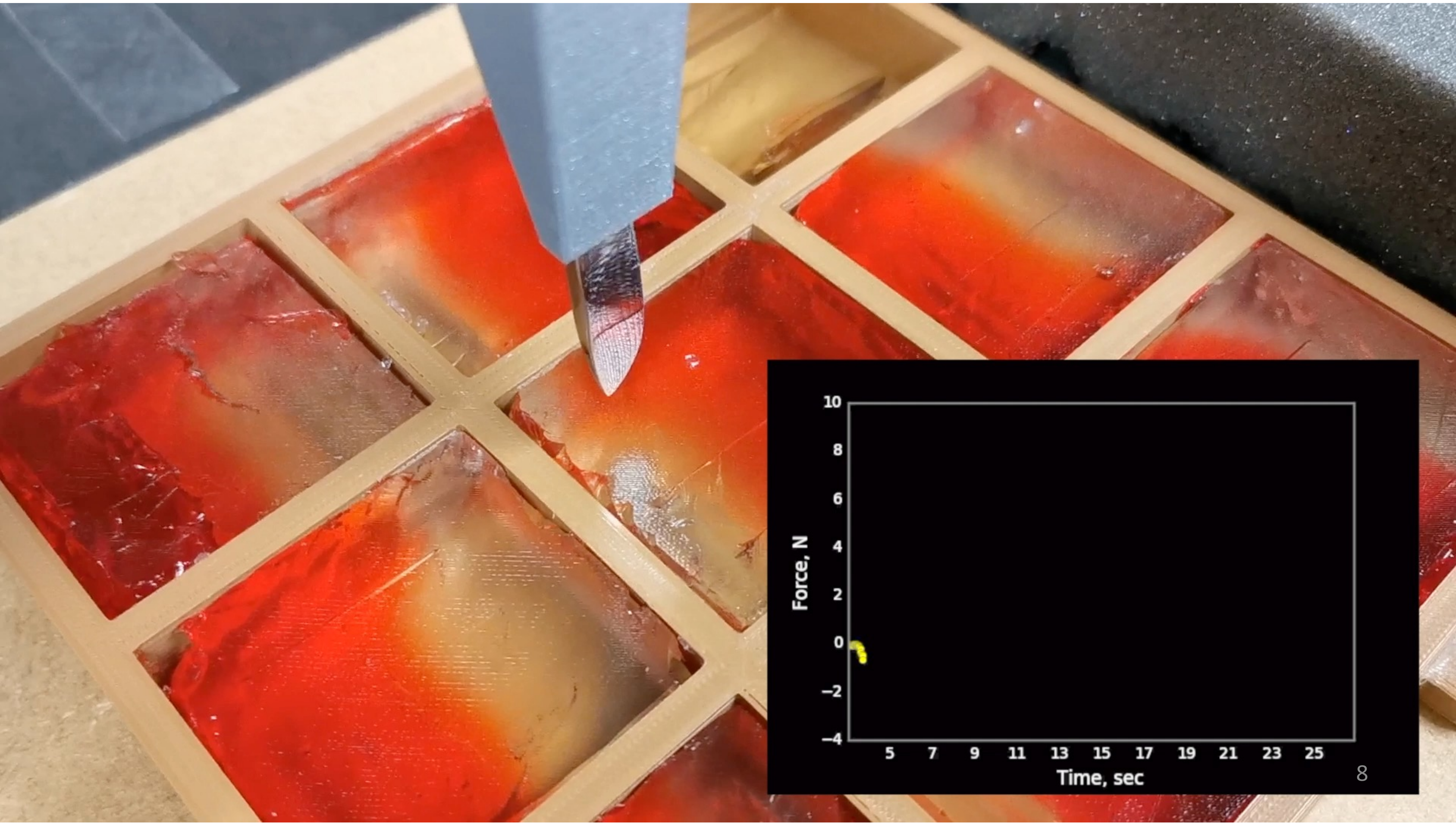
We Do Not Want to Decouple Sensing from Dynamics/Control

Challenges:

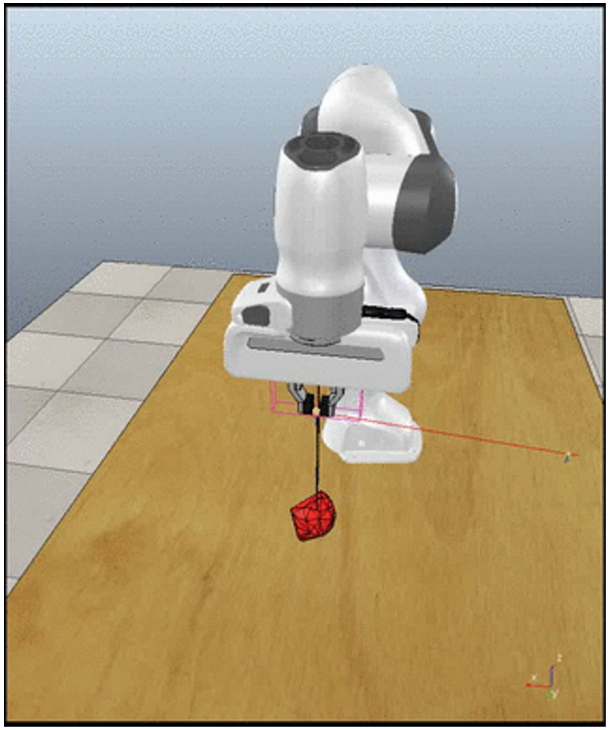
- Path feasibility (kinematic constraints) because path is produced independently.
- Perceptual uncertainty (no closed loop sensing)
- Complex optimisation problem (Planning hierarchies, etc.)

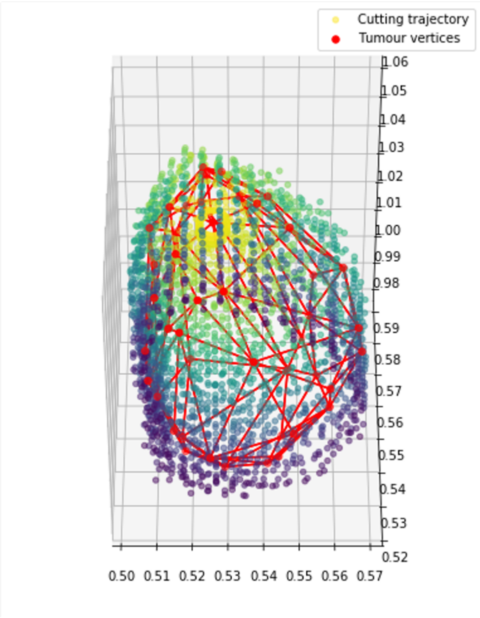


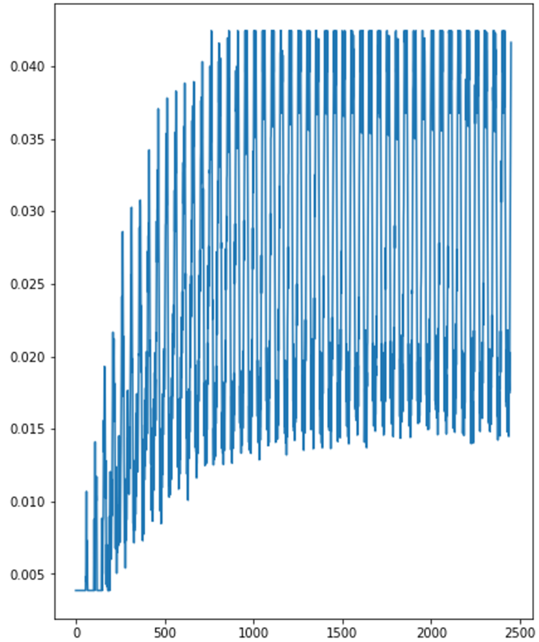
Key Sensory Input: Feel of Tissue



Cutting using Inferred Tactile Response

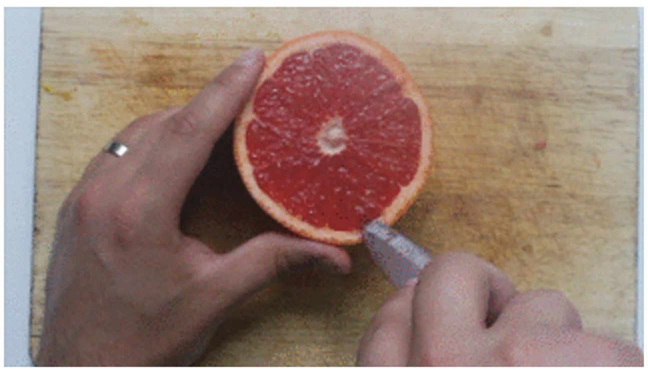


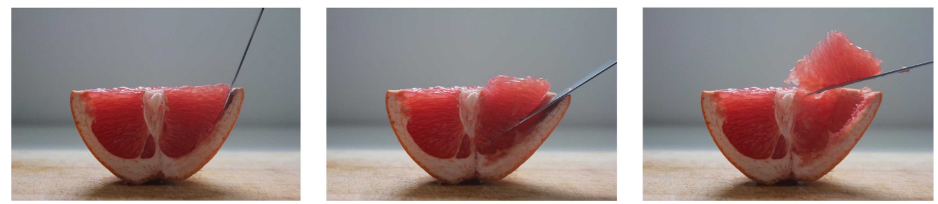


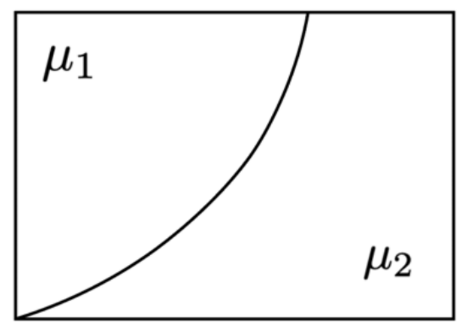


Shape inferred using a controller trying to emulate inferred tactile response.

Would this Work in a Physical Setup?



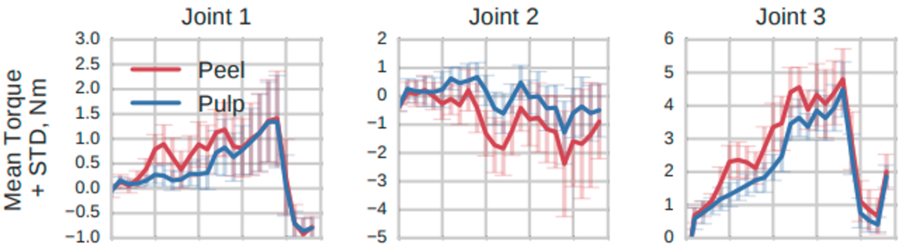


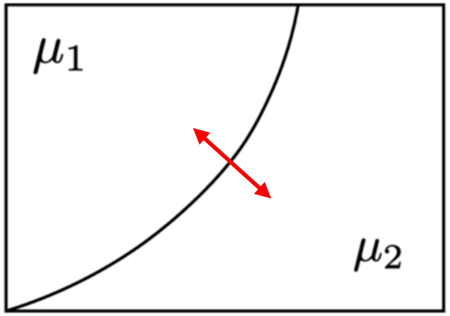


We could try to model the interface between the peel and pulp using FEM, design a cost function giving the dynamics of a knife, and formulate this as an optimal control problem...

… or, we could learn a proportional controller *correction policy to a template* curve model.

Control using the decision boundary of medium classifier as feedback error



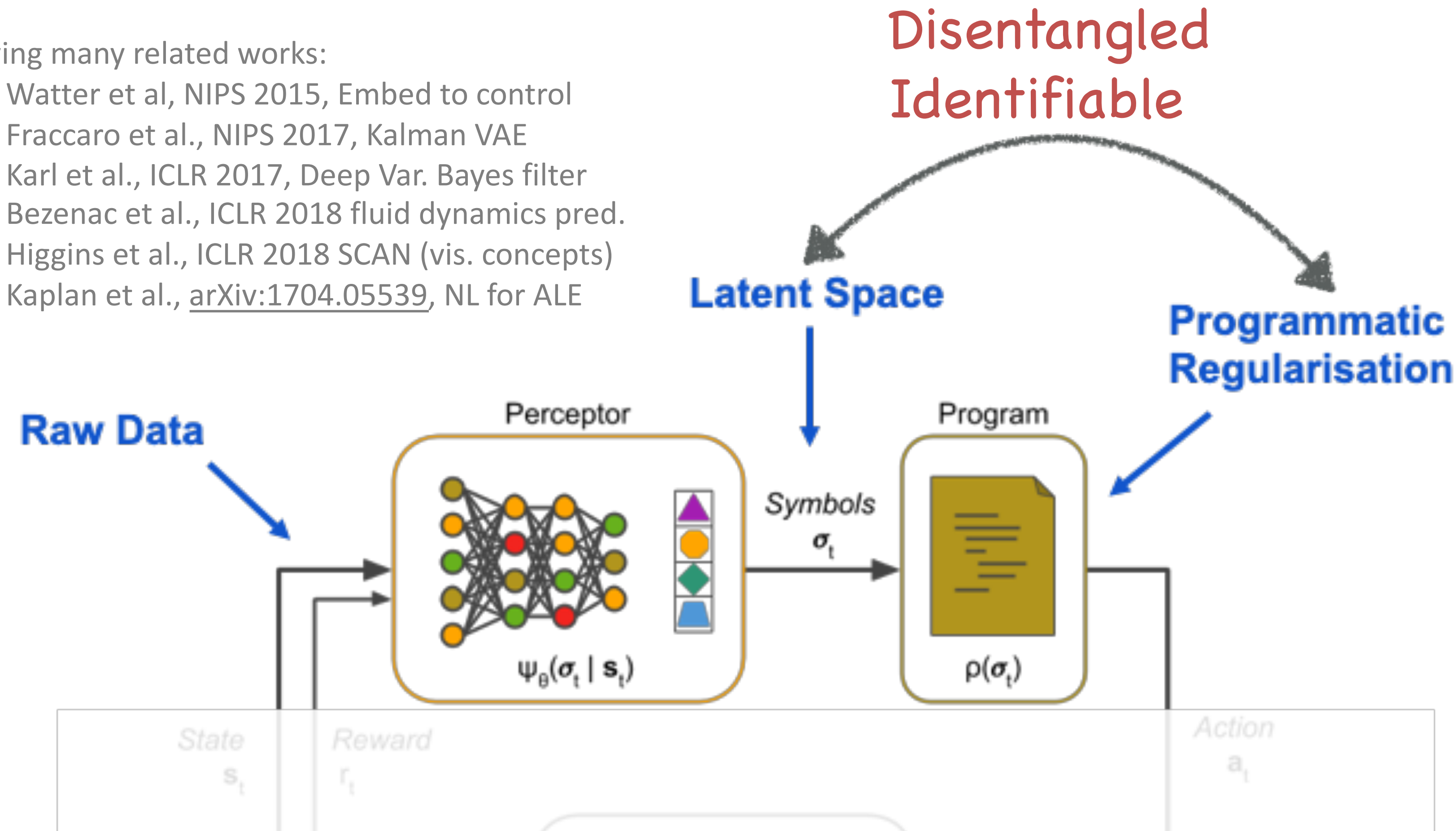


A. Straizys, M. Burke, S. Ramamoorthy, **Surfing on an uncertain edge: Precision cutting of soft tissue using torque-based medium classification**, ICRA 2020. https://arxiv.org/abs/1909.07247

Are We Losing Generality? No!
Can Have State-of-the-art Integrated Schemes

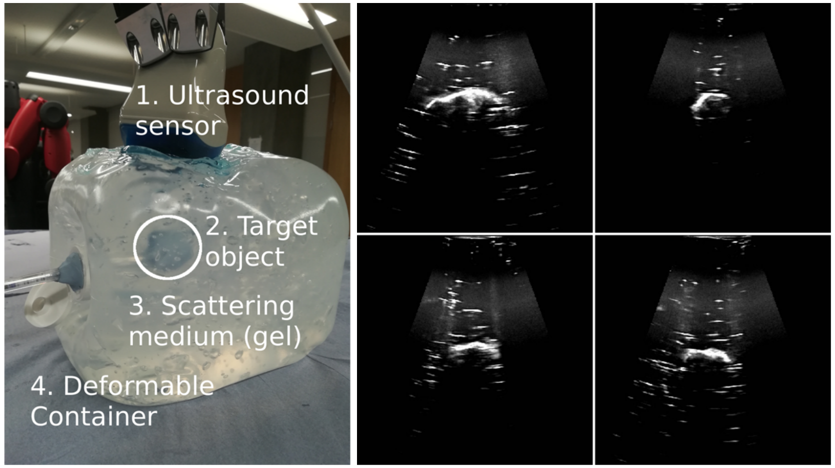
Following many related works:

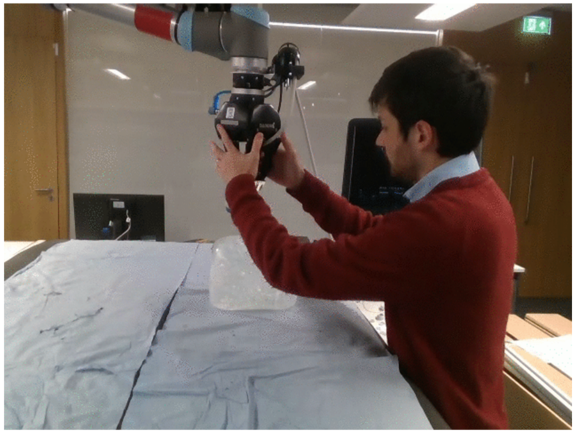
- Watter et al, NIPS 2015, Embed to control
- Fraccaro et al., NIPS 2017, Kalman VAE
- Karl et al., ICLR 2017, Deep Var. Bayes filter
- Bezenac et al., ICLR 2018 fluid dynamics pred.
- Higgins et al., ICLR 2018 SCAN (vis. concepts)
- Kaplan et al., arXiv:1704.05539, NL for ALE



S.V. Penkov, S. Ramamoorthy, **Learning programmatically structured representations with perceptor gradients**, In Proc. *International Conference on Learning Representations (ICLR)*, 2019.

Learning Ultrasound Scanning: Temporal Order as Supervisory Signal

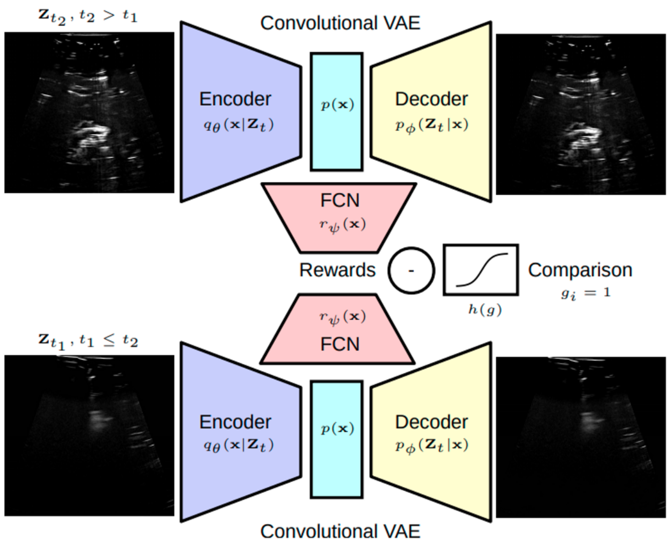


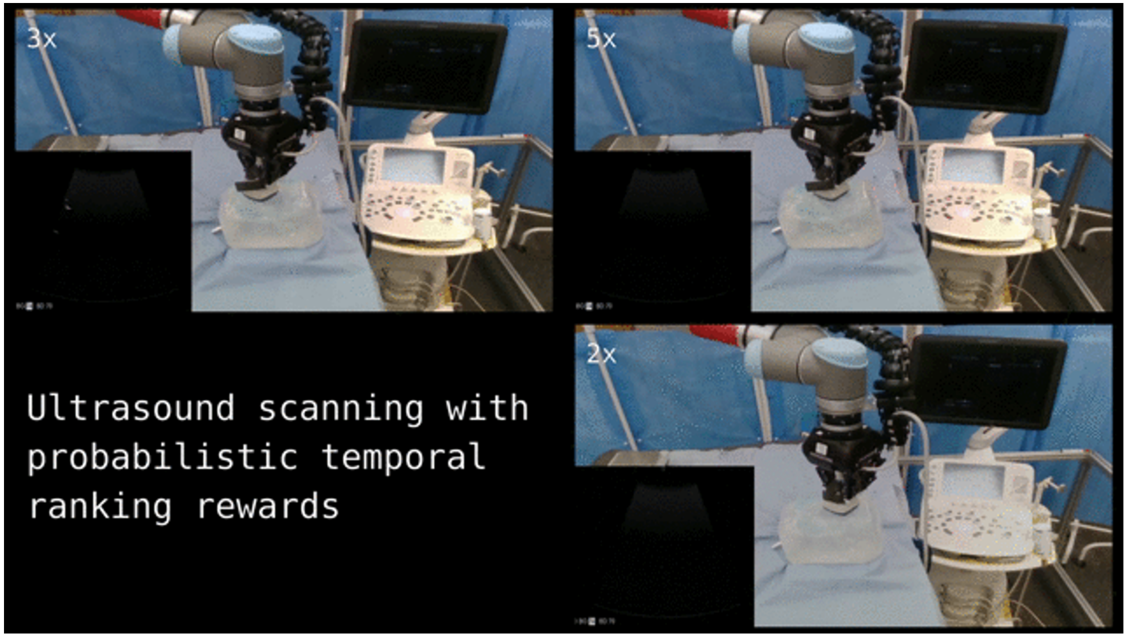


https://sites.google.com/view/ultrasound-scanner

M. Burke, K. Lu, D. Angelov, A. Straižys, C. Innes, K. Subr, S. Ramamoorthy, **Learning rewards for robotic ultrasound scanning using probabilistic temporal ranking**. https://arxiv.org/abs/2002.01240

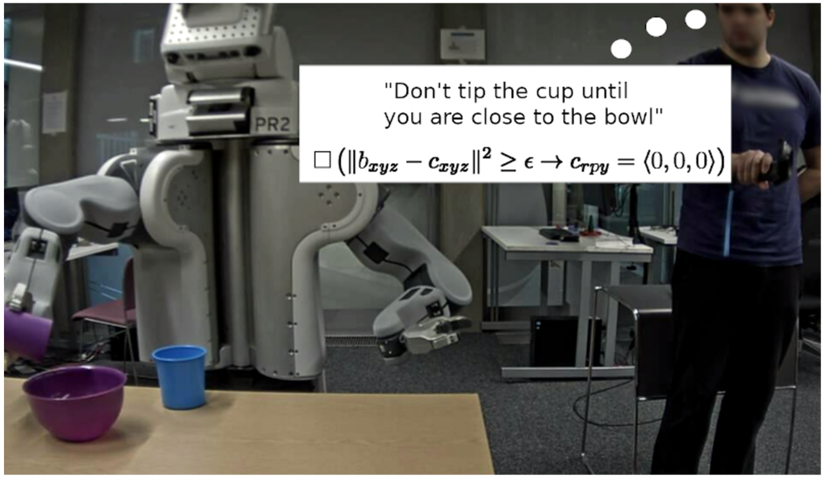
Autonomous Ultrasound Scanning



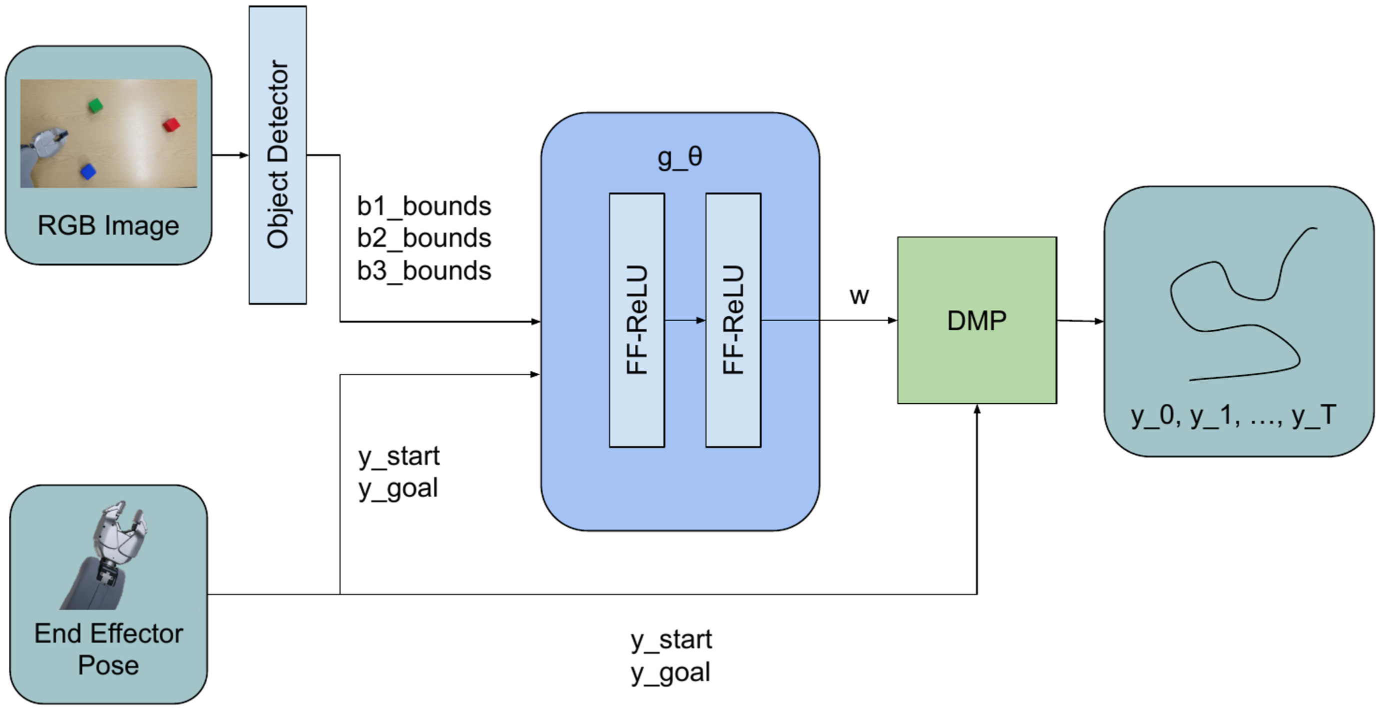


Towards *Safe* Synthesis

Incremental Elaboration of Demonstrated Tasks



C. Innes, S. Ramamoorthy, **Elaborating on learned demonstrations with temporal logic specifications**, *Robotics: Science and Systems* (R:SS), 2020.

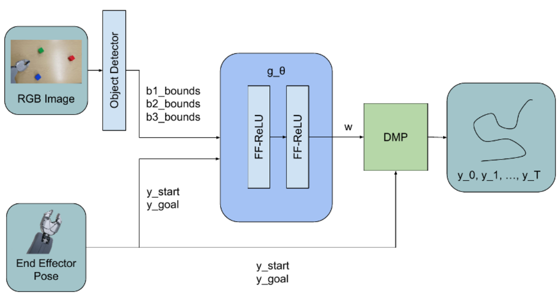


Dynamic Motion Primitive (DMP):

$$\ddot{y} = \alpha_y(\beta_y(y_{goal} - y) - \dot{y}) + f(x)$$
$$\dot{x} = -\alpha_x x$$

$$f(x) = \frac{\sum_{i=0}^{N} w_i \psi_i(x)}{\sum_{i=0}^{N} \psi_i(x)} x(y_{goal} - y_{start})$$

How to Incorporate Specifications?



Don’t Tip the Cup Until you are Close to the Bowl

Elaborating on Demonstrations with Linear Temporal Logic (LTL)

The first thing must remain true **until** the second thing becomes true

$$\varphi := p \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathcal{N}\varphi \mid \Box\varphi \mid \Diamond\varphi \mid \varphi_1 \mathcal{U} \varphi_2$$

This will **always** be true

Eventually, this will be true

Training DMPs with Combined Losses

Imitation / Behaviour Cloning Loss

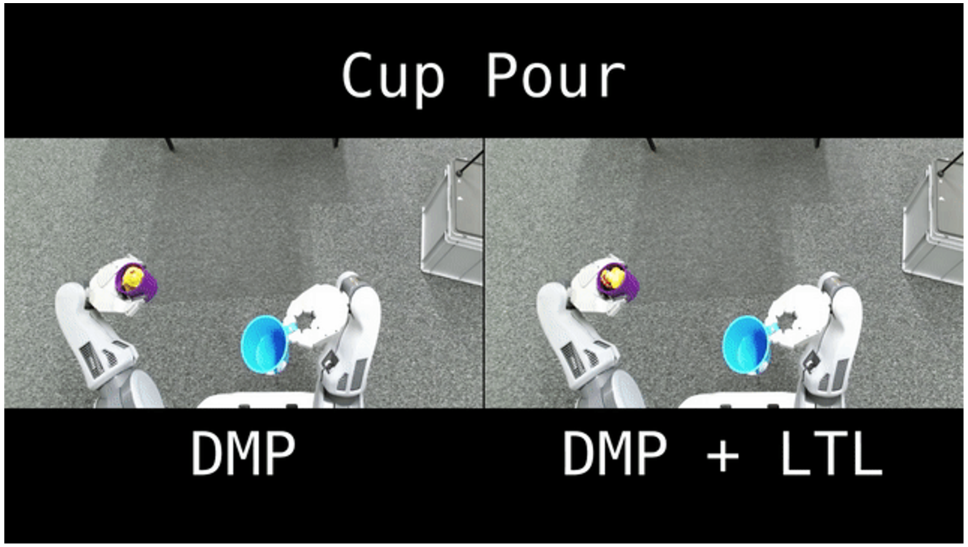
$$\mathcal{L}_d(\theta, x_i, y_i) = \frac{1}{T}\sum_{t=0}^{T} \|\mathrm{DMP}_{\theta,x_i}(t) - y_{i,t}\|^2$$

LTL Constraint Loss
(With Adversarial Examples)

$$z_i^* = \operatorname*{argmin}_z \mathcal{L}_c(\neg\varphi, 0)(\theta, z, y_i)$$

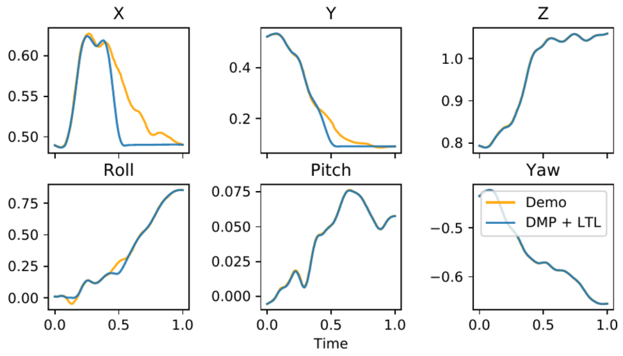
$$\mathcal{L}_{full}(\theta, D, \varphi) = \frac{1}{M}\sum_{i=0}^{M} \mathcal{L}_d(\theta, x_i, y_i) + \eta \mathcal{L}_c(\varphi, 0)(\theta, z_i^*, y_i)$$

Refining Demonstrated Movement

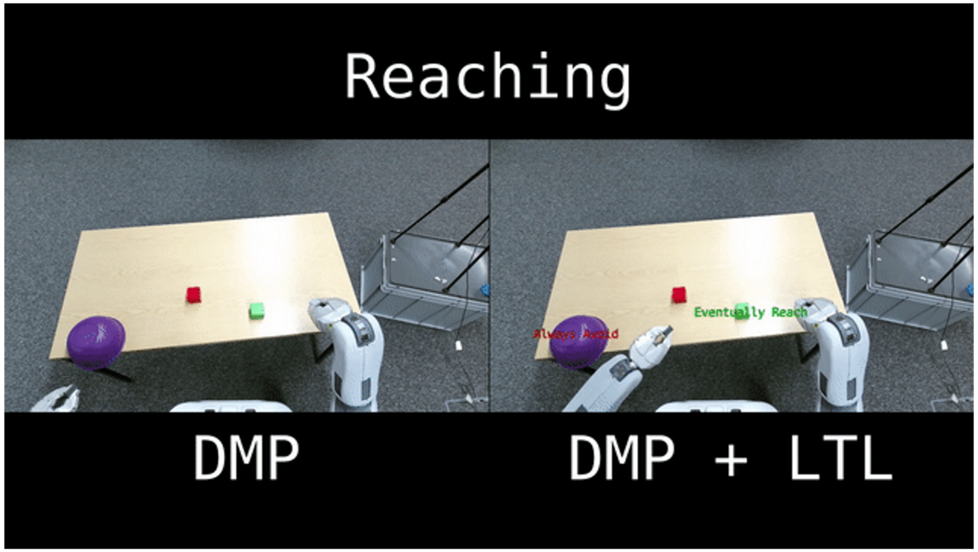


"Also: Don't tip the cup until you're near the container"

$$\square(\|p_{xyz} - x_{i,2}\|^2 \geq 0.1 \ \wedge \ p_z \geq x_{i,2,z}) \implies (\langle 0, 0, -1.0\rangle \leq p_{rpy} \leq \langle 0.2, 0.2, 0.0\rangle)$$



Elaborating on Additional Goals



"Also: Visit the green cube at some point while avoiding the purple bowl"

$$(\Box \|p_{xyz} - x_{i,2}\|^2 \geq 0.2) \wedge (\Diamond p_{xyz} = x_{i,3})$$

Trustworthy Autonomous Systems

The UKRI TAS Programme

The TAS Hub is funded as part of the Strategic Priorities Fund (SPF) which funds multi- and interdisciplinary research across 34 themes in response to strategic priorities and opportunities.

Total Funding

£33m over 4 years

Universities

20+

Industry Partners

100+

Funding

Hub: £11.7m
Nodes : £3m each

Researchers

130+

Disciplines

10+

World's largest research programme in Trustworthy AI and Autonomous Systems

UKRI Research Node on TAS Governance & Regulation

42 months project (Nov 2020 – Apr 2024)

- Two phases:
  - Phase 1 [24 months]: Develop frameworks and smaller demonstrators (6 PDRAs)
  - Phase 2 [18 months]: Large case studies with partners; regulatory sandbox (4 PDRAs)

- Four sub-teams:
  - Legal and social studies
  - TAS Modelling: Causality, explainability, accountability, responsibility
  - Computational tools: formal methods, NLP
  - HCI and Design ethnography

- Project Partners: Adelard, Altran UK, BAE Systems, Civil Aviation Authority (CAA IC), Craft Prospect, Digital Health and Care Inst., D-RisQ, DSTL, Ethical Intelligence, Imandra, Legal and General's ACRC, Imandra, Microsoft, MSC Software, NASA Ames, NPL, NVIDIA, Optos, SICSA, Thales, Vector Four

Concluding Remarks

- Autonomy with learning enabled robots represents many opportunities
- Correspondingly, many fundamental AI challenges:
  - Physics informed machine learning
  - Safety and specification gaps
- Some underexplored issues of immediate interest:
  - Learning **dexterous manipulation** from experts, e.g. surgeons
  - Using **multiple sensory modalities *in situ***, ***in vivo,*** from imaging and spectroscopy to haptics
  - Use of automated ***real-time interpretation*** **within the closed-loop** of autonomous robotic behaviour