When to trust a self-driving car...

Prof. Marta Kwiatkowska

Department of Computer Science
University of Oxford

Based on 2018 Milner Award Lecture at the Royal Society
Exciting times for our discipline!

Robots roam everywhere

Credits: That’s Really Possible, Yamaha/SRI
Beyond digital circuits

DNA computing

“Computing with soup”
(The Economist 2012)

Single DNA strands are inputs and outputs
Also nanostructures and nanorobots

Pop quiz, hotshot: what's the square root of 13?
Science Photo Library/Alamy

“Scientists invented AI made from DNA”
Caltech News

[Qian, Winfree, Science 2012]
[Cherry, Qian, Nature 2018]
Chips with everything no more...

- **Hardware** systems (circuits, communications technology) more like **software**
  - e.g. programmable networks
- **Software methodologies** are on the rise
- **We are already** debugging DNA programs!
Deep learning with everything

DeepFace
Closing the Gap to Human-Level Performance in Face Verification

Yaniv Taigman
Ming Yang
Marc'Aurelio Ranzato
Lior Wolf
- 2014

97.35% accuracy
Trained on the largest facial dataset – 4M facial images belonging to more than 4,000 identities.

Google Translate—here shown on a mobile phone—will use deep learning to improve its translations between texts.

Build for voice with Alexa
Learn more
Much excitement about self-driving…

www(bsfilms.me) – Black Sheep Films
Out and about in Oxford....

Day 4
The challenge of autonomous driving

- Complex engineering and AI problem...
- Software at the heart
- Old and new technologies
  - computer vision
  - sensor fusion
  - control
  - prediction
  - planning
- Increasing use of deep learning
  - requiring high quality data
  - powered by GPUs
- Deep science
- Great progress!

Credit: Oxford Robotics Institute

NVIDIA DRIVE PX 2
Would you trust a self-driving car?

We’re looking to learn from people with diverse transportation needs. Here are some of the first riders who are already using our self-driving cars every day.

Ted and Candace

A typical day in Ted and Candace’s household is full of busy activities across both the parents and their four children: Abbi, Brielle, Izzy and Trey. This lively family is now using our self-driving cars to get to work, shuttle four kids to school and juggle everything from the parents’ weekly date night to their children’s soccer practice. They are excited about giving everyone in their home a greater sense of freedom and independence.

Waymo early riders, Tesla, Uber, ...
In the UK FiveAI, Oxbotica, ...
Unwelcome news recently...

Fatal Tesla Crash Raises New Questions About Autopilot System

U.S. Safety Agency Criticizes Tesla Crash Data Release

How can this happen if we have 99.9% accuracy?
An AI safety problem…

• Complex scenarios
  – goals
  – perception
  – autonomy
  – situation awareness
  – context (social, regulatory)
  – trust
  – ethics

• Safety–critical, so guarantees needed

• Should failure occur, accountability needs to be established
It’s about provable guarantees!

- Modelling = rigorous, mathematical abstraction
- Verification = proof that the model satisfies specification
- Synthesis = correct-by-construction model from specification
- Automated = algorithmic, implemented in software
Probabilistic guarantees

- **Stochasticity ever present**
  - randomisation, uncertainty, risk

- **Need quantitative, probabilistic guarantees for:**
  - safety, security, reliability, performance, resource usage, trust, authentication, …

- **Examples**
  - (reliability) “the probability of the car crashing in the next hour is less than 0.001”
  - (energy) “energy usage is below 2000 mA per minute”

- **My focus is on automated, tool-supported methodologies**
  - probabilistic model checker **PRISM**, [www.prismmodelchecker.org](http://www.prismmodelchecker.org)
  - HVC 2016 Award (joint with Dave Parker and Gethin Norman)

- **Applied to a wide range of systems…**
OK, but what is probabilistic verification good for?
Case study: Cardiac pacemaker

• **How it works**
  – *reads* electrical signals through sensors in the right atrium and right ventricle
  – monitors the **timing** of heart beats and local electrical activity
  – generates *artificial* pacing signal as necessary

• **Safety—critical system!**
• **The guarantee**
  • *(basic safety)* maintain 60–100 beats per minute

  – **Killed by code: FDA recalls 23 defective pacemaker devices because of adverse health consequences or death, six likely caused by software defects (2010)*
Modelling framework

Model the pacemaker and the heart, compose and verify

Quantitative verification of implantable cardiac pacemakers over hybrid heart models. Chen et al, Information and Computation 2014
module VRP

s_vrp: [0..2] init 0;

// Invariants for clock t_vrp
Invariant
(s_vrp = 2 => (t_vrp <= TVRP)) &
(s_vrp = 1 => (t_vrp <= 0));

endInvariant

[Vget] (s_vrp = 0) -> (s_vrp' = 1) & (t_vrp'=0);

[Vt] (s_vrp = 0) -> (s_vrp' = 2) & (t_vrp' = 0);
module VRP

s_vrp: [0..2] init 0;
\[t_vrp\] : clock;

// Invariants for clock \[t_vrp\]

\textit{Invariant}
\begin{align*}
    & (s_{\text{vrp}} = 2) \Rightarrow (t_{\text{vrp}} \leq TVRP) \& \\
    & (s_{\text{vrp}} = 1) \Rightarrow (t_{\text{vrp}} \leq 0)
\end{align*}

\textit{End invariant}

\begin{align*}
[V_{get}] & (s_{\text{vrp}} = 0) \rightarrow (s_{\text{vrp}}' = 1) \& (t_{\text{vrp}}' = 0) \\
[VP] & (s_{\text{vrp}} = 0) \rightarrow (s_{\text{vrp}}' = 2) \& (t_{\text{vrp}}' = 0)
\end{align*}
Pacemaker verification

**Basic guarantees**
- *(basic safety)* maintain 60–100 beats per minute
- *(energy usage)* detailed analysis, plotted against timing parameters of the pacemaker

**Advanced guarantees**
- rate-adaptive pacemaker, for patients with chronotropic deficiency
- *(advanced safety)* adapt the rate to exercise and stress levels
- *in silico* testing

*Closed-Loop Quantitative Verification of Rate-Adaptive Pacemakers.* Paoletti *et al*, ACM Transactions on Cyber-Physical Systems 2018
Synthetic ECG: healthy heart
Bradycardia (slow heart rate)
Bradycardia heart, paced
Parameter synthesis for pacemakers

- Can we adapt the pacing rate to patient’s ECG to
  - minimise energy usage?
  - maximise cardiac output?
  - explore trade offs?

- The guarantee
  - (optimal timing delay synthesis): find values for timing delays that optimise a given objective, adapted to patient’s ECG

- Significant improvement over default values

*Synthesising robust and optimal parameters for cardiac pacemakers using symbolic and evolutionary computation techniques*. Kwiatkowska *et al*, HSB’16
Trade offs in optimal delay synthesis
Case study: ECG biometrics

- Biometrics increasing in popularity
  - are they secure?

- Nymi band
  - ECG used as a biometric identifier
  - biometric template created first
  - compared with real ECG signal

- Proposed uses
  - for access into buildings and restricted spaces
  - for payment
  - etc

Broken Hearted: How to Attack ECG Biometrics, Ebertz et al., In Proc NDSS 2017
Attack on ECG biometrics

- We use synthetic ECGs to impersonate a user
  - build model from data, 41 volunteers
  - inject synthetic signals to break authentication
  - 80% success rate

- Results
  - serious weakness
  - countermeasures needed

- Modelling essential, good for attacks…
Case study: Transferability of attack

• Beware your fitness tracker!
• How easy it is to predict attacks when collecting data from different sources
  – ECG
  – eye movements
  – mouse movements
  – touchscreen dynamics
  – gait
  – etc

• Human study
  – easy for eye movements
  – ECG more chaotic

*When your fitness tracker betrays you*, Ebertz et al., In *Proc S&P 2018*
Case study: DNA origami tiles

- DNA origami tiles: molecular breadboard [Turberfield lab]
- Computation performed by molecular walkers on ‘tracks’
- Build an abstract predictive Markov chain model

DNA walker circuits

- Computing with DNA walkers

- Branching tracks laid out on DNA origami tile, any Boolean function
- The guarantee? walker rates for guaranteed reliability level

Dimer origami
Prediction of dimer origami folding

- Model $2^{76}$ states
- Gillespie simulation
- Remarkable predictive ability

Modified tile

Observed shape

Predicted

N=59 (44%)

N=223 (52%)
Back to the challenge of autonomous driving...

- Things that can go wrong in perception software
  - sensor failure
  - object detection failure

- Machine learning software
  - not clear how it works
  - does not offer guarantees

- Opportunities for the keen scientist!

Lidar image, Credit: Oxford Robotics Institute
Why worry about safety of self-driving?

- Deep neural networks are unstable wrt adversarial perturbations
  - Nexar Traffic Light Challenge: red light classified as green with 68%/95%/78% confidence after one pixel change
- Can reduce to 0% accuracy: can we compute guarantees for neural networks?

German traffic sign benchmark...

<table>
<thead>
<tr>
<th>Stop</th>
<th>30m speed limit</th>
<th>80m speed limit</th>
<th>30m speed limit</th>
<th>Go right</th>
<th>Go straight</th>
</tr>
</thead>
</table>

Confidence: 0.999964 0.99

Aren’t these artificial?

Real traffic signs in Alaska!

Need to consider *physical* attacks, not only digital…
Can also attack 3D deep learning…

...reduce accuracy to 0% after occlusion of 6.5% of the occupied input space, targeting the critical set

**New challenge: verification for ML**

- **What’s different about machine learning?**
  - black box, lacks interpretability
  - programming by pattern matching, not logic
  - corner cases are unseen examples, not missed conditions
  - data quality and coverage crucial
  - accuracy can be misleading

- **Why is ML difficult to verify?**
  - foundations of ML not well understood, mix of logic and real valued functions
  - training obscure, not clear how to choose the training method
  - dependence on choice of loss functions and optimisation
  - scalability an issue

- **Need synthesis, not just verification…**
Guarantees for deep learning!

- Prove that no adversarial examples exist in a neighbourhood around an input
- Compute lower and upper bounds on maximal safety radius

A Game-Based Approximate Verification of Deep Neural Networks with Provable Guarantees, Wu et al, CoRR abs/1807.03571, 2018.
Probabilistic guarantees

• Requiring that no adversarial examples exist too strict!

• Need to **probabilistic guarantees**: probability that local perturbations result in predictions that are close to original

• Taking account of the **learning** process

• Bayesian neural networks have **prior** on weights
  – account for noise, uncertainty, etc
  – return an uncertainty measure along with the output

• Need to compute posterior probability
  – often **intractable**
  – can we do better?
Statistical robustness guarantees

- Work with Bayesian neural networks
- Define safety with prob $1 - \varepsilon$
  \[ \text{Prob}(\exists y \in \eta \text{ s.t. } f(x) \neq f(y) \mid D) \leq \varepsilon \]
- i.e. conditioned on training data $D$
- Method: sample the weights, then employ statistical model checking (Massart bounds, sequential test)
  - compare robustness and accuracy trade offs for different inference methods

Robustness comparison

So have we solved the problem?

'I hate them': Locals reportedly are frustrated with Alphabet's self-driving cars

- Alphabet's self-driving cars are said to be annoying their neighbors in Arizona, where Waymo has been testing its vehicles for the last year.
- More than a dozen locals told The Information they they hated the cars, which often struggle to cross a T-intersection near the company's office.
- The anecdotes highlight how challenging it is for self-driving cars, which are programmed to drive conservatively, to handle certain situations.

Self-driving cars should be allowed to mount pavements and break speed limit in emergencies

Published 2:04 PM ET Tue, 28 Aug 2018 | Updated 12:53 PM ET Wed, 29 Aug 2018

Source: Waymo

A Tesla Model S
Trust, ethics, morality and social norms...

- Already merging into traffic proving difficult,
  - what about social subtleties?
  - communication, multi-modal signals?

- Need to reason about
  - trust
  - moral decisions
  - conflict resolution
  - accountability: black box?

- Already developing quantitative verification for trust...

http://www.pbs.org/wgbh/nova/next/tech/robot-morals/
Concluding remarks

• Overview of role of probabilistic modelling, verification and synthesis
  – safety/performance guarantees, prediction, attacks, optimal synthesis, and more

• Much excitement about potential of the developments in AI
  • and exciting opportunities!
• But need to know the limits, also for deep learning
  – rigorous foundations, formal verification, safety assurance

• and social implications
  – overtrust/undertrust in robots
  – ethics of autonomous decision making
  – morality of autonomous behaviour

• Many challenges remain
A tribute to Robin Milner

• From computers to ubiquitous computing, by 2020

“The most profound technologies are those that disappear. They weave themselves into everyday life until they are indistinguishable from it.” (Weiser, 1993)

• “Ubicomp can empower us” (Milner)
• We must keep a live connection between theory and application in computer science

• This lecture is a contribution
  – practical, algorithmic techniques and industrially-relevant tools
Acknowledgements

- My group and collaborators in this work
- Project funding
  - ERC Advanced Grant
  - EPSRC Mobile Autonomy Programme Grant
- See also
  - PRISM www.prismmodelchecker.org
- **New** ERC Advanced Grant FUN2MODEL
  “From FUNction–based TO MOdel–based automated probabilistic reasoning for DEep Learning”
- Postdoctoral and PhD positions
Acknowledgements

• **Contributors** (to PRISM & its underlying theory)
  
  - Aistis Simaitis, Alberto Puggelli, Alessandro Bruni, Alexandru Mereacre, Alistair John Strachan, Andrej Tokarčík, Andrew Hinton, Antonio Pacheco, Archit Taneja, Ashutosh Trivedi, Benoit Barbot, Bruno Lacerda, Carlos Bederian, Charles Harley, Chris Thachuk, Christel Baier, Christian Dehnert, Christian von Essen, Christopher Ziegler, Chunyan Mu, Clemens Wiltsche, Dave Parker, Ernst Moritz Hahn, Frits Dannenberg, Fuzhi Wang, Ganindu Prabhashana, Gethin Norman, Håkan Younes, Holger Hermanns, Hongyang Qu, Jan Křetínský, Jens Katelaan, Jeremy Sproston, Joachim Klein, Joachim Meyer-Kayser, Joost-Pieter Katoen, Kenneth Chan, Klaus Draeger, Kousha Etessami, Lovejeet Singh, Lu Feng, Luca de Alfaro, Marcin Copik, Marco Diciolla, Maria Svorenova, Mark Kattenbelt, Markus Siegle, Marta Kwiatkowska, Mateusz Ujma, Maximilian Probst, Mihalis Yannakakis, Mike Arthur, Milan Ceska, Moshe Vardi, Muhammad Omer Saeed, Nick Hawes, Nicola Paoletti, Nicolas Basset, Nicolas Del Piano, Nishan Kamaleson, Paolo Ballarini, Pedro D'Argenio, Qixia Yuan, Radu Calinescu, Rashid Mehmood, Roberto Segala, Sebastian Vermeiren, Sergio Giro, Steffen Märcker, Stephen Gilmore, Taolue Chen, Tingting Han, Vincent Nimal, Vojtěch Forejt, Xueyi Zou, Yi Zhang, Zak Cohen, ...  
  
  (and many more collaborators on case studies & projects)

• **Project funders**