Heidy A. Khlaaf

http://heidykhlaaf.com

EDUCATION

University College London **2013 - 2017**
PhD: Computer Science
Advisors: Nir Piterman
Topic(s): Formal Verification, Temporal Logic, and Model Checking of Infinite-State Systems

Florida State University **2012**
Bachelor of Science: Computer Science, Philosophy
Minor: Mathematics
Summa cum laude, Honors, Phi Beta Kappa; **3.9 GPA**

EXPERIENCE & RESEARCH

AI Now **Remote June 2024 - Present**
Chief AI Scientist

- Leading the research and assessment of safety and security of AI use within military systems and armaments.

Trail of Bits **Remote Nov 2022 - March 2024**
Engineering Director - ML Assurance

- Founded and led the ML assurance practice, focusing on the security, risk assessment, and auditing of AI & ML based systems.
- Led the cyber evaluations for the UK Government's AI Safety Institute and carried out an impact assessment on the global proliferation of AI cyber capabilities on national security (presented at the UK AI Safety Summit).
- Oversaw and contributed to the discovery of novel vulnerabilities identified within the ML pipeline including LeftoverLocals and Sleepy Pickle.

Zipline **Remote Feb 2021 - Feb 2022**
Lead Software Safety Engineer

- Led Zipline's Software Safety strategy to enable the assurance of autonomous UAVs and ensuring their safe deployment within the US airspace.
- Led cross-functional system hazard analysis and risk assessment focused on identifying risks and failures introduced by state-of-the-art Detect and Avoid ML-based system.
- Authored software certification plans (SDP, SVP, SCMP, SQAP) to be submitted under the FAA's Means of Compliance for Software (e.g., DO-178C).
- Devised novel software certification frameworks with Legal and Regulatory to permit the assurance of disruptive technologies (i.e., ML models, BVLOS Drones) that lack regulatory precedent.

OpenAI **Remote Dec 2020 - May 2021**
Senior Systems Safety Engineer - Contractor

- Led Copilot's and Codex's Safety and Programming Language analysis and formulated corresponding system and organizational mitigations.
- Developed a framework for measuring code generation models' performance based on the the complexity of natural language specifications.

- Introduced and devised a cross-functional hazard analysis and risk assessment framework focused on identifying risk factors that address novel safety, societal, and security risks imposed by state-of-the-art ML models (Codex & GPT-3).

Adelard LLP **London, UK June 2017 - Nov 2020**
Senior Consultant

- Evaluating, specifying, and verifying safety-critical systems and dependable computing applications using various techniques including: formal methods, model checking, and a variety of other static analysis techniques and tools.
- Managed and technically led projects with various clientele including: EDF Energy, Hinkley Point C, Office for Nuclear Regulation, Department for Transport, Centre for the Protection of National Infrastructure (MI5), DARPA, and many others.
- Led safety and security audits that contribute to the assurance of clients' projects (e.g., IEC 61508) in Energy and Industrial Automation, and assisted with hazard analysis, provided independent advice, and reviewed and contributed to safety and security assurance cases.
- Identified system and software risks and failure modes and devised mitigation strategies to achieve the required Safety Integrity Level (SIL) and reliability requirements.
- Developed and contributed to government policy and regulatory frameworks to enable the assurance of disruptive technologies such as artificial intelligence and machine learning to be deployed within safety-critical systems.
- Constructed safety cases using the Claims, Arguments, Evidence (CAE) framework for safety and security related applications and their development.
- Extended the CAE framework to support novel technologies (e.g., ML) that fall outside of regulatory scope in order to scrutinize the safety justification of their use within safety-critical systems.

Sample work:

- As technical lead, directed and carried out the independent software verification and validation of the Safety Automation System (Class 2) in the Hinkley Point C nuclear power plant based on an architectural and requirements analysis. This included determining the analytical reliability tradeoffs and the feasibility of V&V strategies, tools, and technologies required to ensure Class 2 regulatory compliance.
- Managed and technically led the UK Nuclear Consortium's (CINIF) IEC 61508 regulatory compliance strategy, including the establishment of auditing requirements and scope, supporting tools and processes, and training strategies to scale and standardize accepted best practices among nuclear stakeholders.
- As technical lead, advised the Office for Nuclear Regulation on the suitability of existing UK nuclear regulation to the application of AI and ML in operations affecting nuclear material, and produced novel regulatory approaches that would enable the safety assurance of AI within nuclear power plants.
- Under CPNI and the Department for Transport, investigated current autonomous vehicles' system development approaches (using PAS 11281 and UL 4600) and identified their assurance gaps and corresponding mitigation strategies that included: novel safety assurance approaches, new ML verification analysis techniques using formal verification, static analysis, and simulation, and an evaluation of defence in depth.

Amazon Web Services **New York, NY June 2015 - Sep 2015**
Automated Reasoning Group - Research Scientist

- Analyzed the application of static analysis methods to resolve a wide variety of SSL certification validation bugs which are pervasive in Amazon's EC2 Java client library, Elastic Load Balancing API Tools, and Amazon Flexible Payments SDK.

Microsoft Research **Cambridge, UK Sep 2013 - Sep 2014**
Formal Methods Research Consultant

- Conducted further research and development to extend the functionality and applicability of the Temporal Logic Verifier **T2** to incorporate strictly more expressive logics such as Fair-CTL and CTL*.

Microsoft Research **Cambridge, UK Jan 2013 - April 2013**
Programming Languages Research Intern

- Discovered how procedure summarization, precondition synthesis, and traditional bottom up approaches complement each other to improve the performance and applicability of novel Computation Tree Logic verification tools.

Microsoft Research **Cambridge, UK May 2012 - Aug 2012**
Programming Languages Research Intern

- Encoded temporal property verification as program analysis task. Produced an encoding which, with the use of recursion and nondeterminism, enables off-the-shelf program analysis tools to naturally perform the reasoning necessary for proving temporal properties in T2.

Florida State University **Tallahassee, FL Sep 2010-Aug 2012**
Research Assistant

- Assisted in the exploitation of parallelism found within functional programming in order to construct an intrinsically parallel language which exhibits intuitive parallel syntax.
- Created a statically-typed functional language that integrates seamlessly with C/C++. The language will have a functional declarative style, will be highly efficient to translate and execute, provides explicit and implicit parallel constructs, list comprehensions, and pattern matching.

PUBLICATIONS

"LeftoverLocals: Listening to LLM Responses Through Leaked GPU Local Memory"
T. Sorensen*, H. Khlaaf*. arXiv:2401.16603 [cs], January 2024.

"Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems"
H. Khlaaf*. Trail of Bits, March 2023.

"A Hazard Analysis Framework for Code Synthesis Large Language Models"
H. Khlaaf*, P. Mishkin*, J. Achiam, G. Krueger, M. Brundage. arXiv:2207.14157 [cs], July 2022.

"Evaluating Large Language Models Trained on Code"
M. Chen, et al. arXiv:2107.03374 [cs], July 2021a.

"97 Things Every SRE Should Know" edited by **E. Stolarsky and J. Woo**. O'Reilly Media Inc., November 2020.

"Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims"

M. Brundage, et al. *59 co-authors from 29 organisations, including: Open AI, Leverhulme Centre for the Future of Intelligence, University of Oxford, Partnership on AI, Adelard, Mila, Google Brain, and many others.* April, 2020.

"Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy"
R. Bloomfield* and H. Khlaaf* with P. Ryan Conmy, G. Fletcher. IEEE Computer, 52(9): 82-89 (2019).

"The Past, Present, and Future(s): Verifying Temporal Software Properties"
H. Khlaaf*. *University College London, Department of Computer Science*, PhD Dissertation, April 2018, London, UK.

"Verifying Increasingly Expressive Temporal Logics for Infinite-State Systems"
H. Khlaaf* with B. Cook and N. Piterman. *Journal of ACM*, 64, 2, Article 15 (May 2017), 39 pages.

"T2: Temporal Property Verification"
M. Brockschmidt* and H. Khlaaf* with B. Cook and N. Piterman. *Tools and Algorithms for the Construction and Analysis of Systems*, Eindhoven, Netherlands, 2016.

"On Automation of CTL* Verification for Infinite-State Systems"
H. Khlaaf* with B. Cook and N. Piterman. *Computer Aided Verification*, San Francisco, USA, 2015.

"Fairness for Infinite-State Systems"
H. Khlaaf* with B. Cook and N. Piterman. *Tools and Algorithms for the Construction and Analysis of Systems*, London, UK, 2015.

"Faster Temporal Reasoning for Infinite-State Programs"
H. Khlaaf* with B. Cook and N. Piterman. *Formal Methods in Computer-Aided Design*, Lausanne, Switzerland, 2014.

COMMUNITY

Working Expert Groups

- *UN Secretary-General's AI Advisory Body* *March 2024 - Present*
  Network of Experts for the UN Secretary-General's High-Level Advisory Body on Artificial Intelligence.
- *British Standards Institute – ART/1 Artificial Intelligence* *Aug 2022 - Present*
  Under the direction of SPSC is the responsibility to mirror the work of ISO/IEC/JTC 1/SC 42 as well as produce independent national portfolio of work.
- *World Economic Forum & SRI* *March 2021 - Present*
  Shaping the Future of Technology Governance: Artificial Intelligence and Machine Learning
- *Federal Aviation Authority (supporting Zipline)* *June 2021 - Feb 2022*
  BVLOS Aviation Rulemaking Committee (Safety)

Program Committee

ACM Conference on Fairness, Accountability, and Transparency (FAccT)	*2023-2024*
PLDI Student Research Competition Judge	*June 2021*
DebugML, ICLR	*March 2019*
Principles of Programming Languages AE	*October 2016*

Computer-Aided Verification AE	*May 2016*
Tiny Transactions on Computer Science (V. IV)	*Jan 2016*
Tiny Transactions on Computer Science (V. II)	*March 2013*

Sub-Review Committee

Tools and Algorithms for the Construction and Analysis of Systems	*Nov 2015*
International Conference on Computer-Aided Verification	*March 2015*
International Conference on Computer-Aided Verification	*February 2014*
International Conference on Computer-Aided Verification	*February 2013*
Formal Methods in Computer-Aided Design	*July 2012*

Program Chair

Tiny Transactions on Computer Science (V. III)	*May 2014 - May 2015*

Activities & Services

UCL - Athena Swan PhD Student Representative	*2016-2017*
UCL - PPLV PhD Student Representative	*2015-2017*
UCL - PhD Student Representative	*2013-2015*
Upsilon Pi Epsilon - Florida State University Chapter President	*2012*
ACM - Florida State University Chapter Undergraduate Vice President	*2012*

Professional Memberships

BCS Professional member (MBCS)
Association for Computing Machinery
Phi Beta Kappa

TEACHING

University College London **London, UK September 2016-Present**

Teaching Assistant

- *COMP204P: Systems Engineering I* Fall 2016
- *COMP205P: Systems Engineering II* Spring 2017

Florida State University **Tallahassee, FL August 2011-Dec 2012**

Teaching Assistant (20 hours/week)

- *Instructed recitation sessions, assessed assignments, projects, exams, and held daily office hours to assist students.*

COP4342 Unix Tools	Fall 2012
COP3330 Object Oriented Programming	Spring 2012
COP3330 Object Oriented Programming	Fall 2011
COP3353 Introduction to Unix	Fall 2011

TALKS & WORKSHOPS

Technical

- *APLAS – SPLASH* *July 2021 Chicago, IL*
  Invited Speaker: "AMA with Heidy Khlaaf"
- *Microsoft Research* *Feb 2021 Redmond, WA*
  Invited Speaker: "Auditing Safety-Critical AI systems"
- *Innovating AI Governance* *Dec 2020 Toronto, CA*
  Shaping the Agenda for a Responsible Future - Schwartz Reisman Institute for Technology and Society
  Invited Participant

- *BSI-VdTÜV AI Forum On Auditing AI-Systems* *Oct 2020 Berlin, Germany*
  Invited Speaker: "Auditing safety-critical AI systems"
- *SRE Con* *Oct 2019 Dublin, Ireland*
  Keynote Speaker: "Applicable and Achievable Formal Verification" (∼800 attendees)
- *AI Vulnerabilities Workshop @ AINow/Google* *Aug 2019 NYC, NY*
  Invited Speaker & Participant
- *Trust in AI Development @ Open AI/PAI* *April 2019 San Francisco, CA*
  Invited Speaker: "Assurance Frameworks for Autonomous Safety Critical Systems"
- *Papers We Love @ StrangeLoop* *Sep 2018 St. Louis, Missouri*
  Invited Speaker: "Standards We Love" (∼500 attendees)
- *F# eXchange* *April 2018 London, UK*
  Invited Speaker: "Lessons from F#: From Academic Prototypes to Safety-Critical Systems"
- *Github Constellation* *March 2018 London, UK*
  Invited Speaker: "Determining Software Safety in Critical Systems" (∼350 attendees)
- *University of East London* *Nov 2017 London, UK*
  Invited Speaker: "Verification of Software Systems, Smart Sensors and the Nuclear Industry"
- *Queen Mary University* *March 2017 London, UK*
  Invited Speaker: "Verifying Increasingly Expressive Temporal Logics for Infinite-State Systems"
- *University of Kent* *Dec 2016 Canterbury, UK*
  Invited Speaker: "Verifying Increasingly Expressive Temporal Logics for Infinite-State Systems"
- *TACAS* *April 2016 Eindhoven, Netherlands*
  Speaker: "T2: Temporal Property Verification"
- *Computer Aided Verification* *July 2015 San Francisco, USA*
  Speaker: "On Automation of CTL* Verification for Infinite-State Systems"
- *TACAS* *April 2015 London, UK*
  Speaker: "Fairness for Infinite-State Systems"
- *University of Leicester* *March 2015 Leicester, UK*
  Invited Speaker: "Verifying Fairness for Infinite-State Systems"
- *Formal Methods in Computer-Aided Design* *Oct 2014 Lausanne, Switzerland*
  Speaker: "Faster Temporal Reasoning for Infinite-State Systems"
- *14th International Workshop on Termination* *July 2014 Vienna, Austria*
  Speaker: "Fairness for Infinite-State Systems"
- *F#unctional Londoners* *March 2013 London, UK*
  Invited Speaker: "T2: A Temporal Property Verifier in F#"

Non-Technical

- *Tech Night LDN* *March 2018 London, UK*
  Invited Panel Speaker: "Diversity in Technology"
- *Microsoft Research* *Dec 2012 Cambridge, UK*
  Keynote Speaker at Think Computer Science 2012

- *Long Road Sixth Form College* *July 2012 Cambridge, UK*
  Invited Guest Speaker

AWARDS AND HONORS

International Conference on CAV - Best Paper Award	*July 2015*
University College London - Research Excellence Studentship	*Sept 2013*
National Science Foundation - Graduate Research Fellowship	*Sept 2013*
Summer School of Marktoberdorf - Attendee	*Aug 2013*
CRA-W/CDC/SIGPLAN Mentoring Workshop at POPL Scholarship	*Jan 2013*
CRA-W/CDC/SIGPLAN Mentoring Workshop at POPL Scholarship	*Jan 2012*
Fall 2011 Bess Ward Honors Thesis Award	*Fall 2011*
Florida State University President's List	*2010 - 2012*
Florida Medallion Scholar	*2008 - 2012*
Florida State University Dean's List	*2008 - 2012*
National SMART Grant recipient	*2008 - 2011*

SKILLS

Languages & Software: C++, C, F#, Perl, LLVM IR, Haskell, C#, Java.
Operating Systems: Adept in Windows, Unix, Linux, and Mac OS.
Other: Fluency in the Arabic Language