Meaningful Human Control of AI Based Systems

Jay Shively

Adaptive Aero Group

Background

- 2016 UN “Meeting of Experts”
- MHC required for “autonomous” weapons
- No Definition of MHC
- 2018 NATO Science and Technology Office (STO) Human Factors and Medicine Panel (HFM) stood up an exploratory team to look at Human Systems aspects of MHC
- 2019 Evolved into 3 year task force – HFM 320

Man who saved the world

In 1983, Stanislav Petrov was in charge of a Soviet nuclear early warning center when there was a report of five American nuclear missiles heading towards the soviet union. Rather than retaliate, Stanislav followed his gut feeling and went against protocol, convincing the armed forces that it was a false alarm. His decision saved the world from a potential devastating nuclear holocaust.

Army of None

- Paul Scharre
- **special operations reconnaissance team leader in the Army's 3rd Ranger Battalion and completed multiple tours to Iraq and Afghanistan**
- "A young girl of maybe five or six headed out of the village and up to our way, two goats in trail. Ostensibly she was just herding goats, but she [was actually] spotting for Taliban fighters." Scharre's unit did not shoot. Yes, it would have been legal, but he argues that it would not have been morally right. But think about it: a robotic sniper following strict algorithms might well have opened fire the second it detected a radio in her hand. As you can see, many unpredictable dilemmas can happen in these scenarios.

Civil applications

Similar concerns arise with civil applications:

- Autonomous cars
- Autonomous air vehicles

AI should NOT make ethical decisions.

NATO STO Human Factors and Medicine - 320

- US/Netherlands co-leads
  - Mark Draper, AFRL
  - Jurrian Van Diggilan, TNO
- Representatives from:
  - France
  - United Kingdon
  - Germany
  - Sweden
  - Finland
  - Netherlands
  - United States
  - Italy
  - Australia

HFM 320

- Workshops
- Invited Speakers
- Two meetings a year for 3 years
- Developed 10 scenarios (including civil) to make concepts more concrete

Definition

The ability of humans to make informed choices in sufficient time to influence AI-based systems in order to enable a desired effect or to prevent an undesired immediate or future effect on the environment.

Guidelines

- Lawfulness
- Responsibility and accountability
- Explainability and traceability
- Reliability
- Governability

Embedded Systems

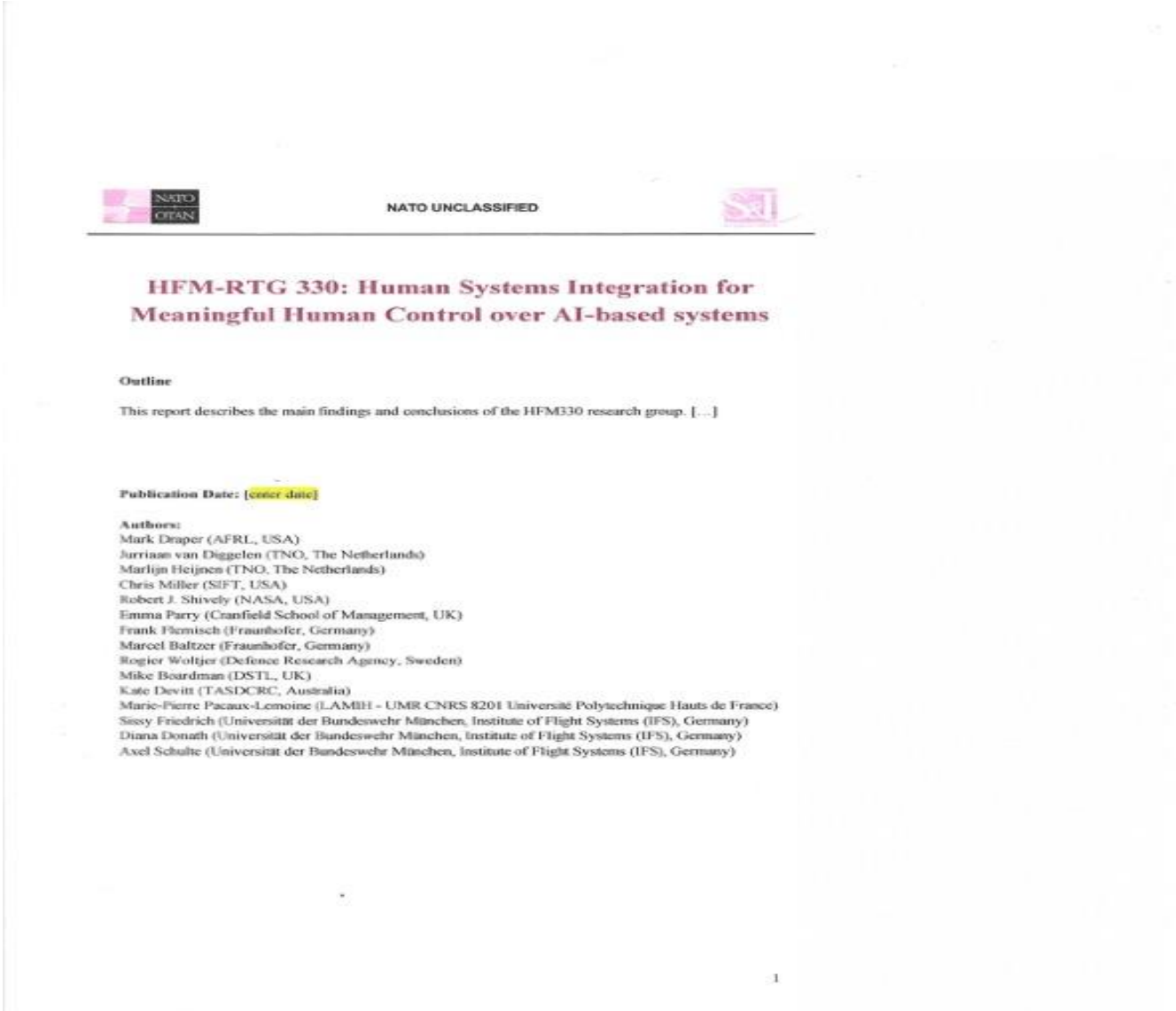
- AI systems can be (and are) nested in ways that make it difficult to know how they have affected the system

- Especially true for perception
  - If the AI directs where a system “looks” or interprets/processes raw data before presenting to a human – then all analysis and decisions downstream are already “biased” by the AI

Working Agreements/Advance Directives

- Real-time MHC is rendered impossible by incident timelines
- Working agreements or Advance Directives can allow human control without being “in” the loop
- If conditions A, B, C exist, the execute X, Y, Z
- If the DAA system detects an aircraft on a course that will result in an unacceptable point of closest approach (PCA) AND no action has been taken AND the PCA will occur within 20 seconds, then perform evasive maneuver

Technical Report due out soon…



NATO UNCLASSIFIED

HFM-RTG 330: Human Systems Integration for Meaningful Human Control over AI-based systems

Outline

This report describes the main findings and conclusions of the HFM330 research group. […]

Publication Date: [enter date]

Authors:
Mark Draper (AFRL, USA)
Jurriaan van Diggelen (TNO, The Netherlands)
Marlijn Heijnen (TNO, The Netherlands)
Chris Miller (SIFT, USA)
Robert J. Shively (NASA, USA)
Emma Parry (Cranfield School of Management, UK)
Frank Flemisch (Fraunhofer, Germany)
Marcel Baltzer (Fraunhofer, Germany)
Rogier Woltjer (Defence Research Agency, Sweden)
Mike Boardman (DSTL, UK)
Kate Devitt (TASDCRC, Australia)
Marie-Pierre Pacaux-Lemoine (LAMIH - UMR CNRS 8201 Université Polytechnique Hauts de France)
Sissy Friedrich (Universität der Bundeswehr München, Institute of Flight Systems (IFS), Germany)
Diana Donath (Universität der Bundeswehr München, Institute of Flight Systems (IFS), Germany)
Axel Schulte (Universität der Bundeswehr München, Institute of Flight Systems (IFS), Germany)

Today?

- Be aware of MHC
  - Human Factors (MANPRINT)
- Examine explainability and traceability of decisions
- Map AI within systems
- Develop scenarios (worst case) and trace decision making