Toward Calibrating Trust in Autonomy

Glenn Taylor
glenn@soartech.com

NDIA Human Systems Conference
June 2022

SOARTECH

Modeling human reasoning.
Enhancing human performance.

This information is approved for release

Definitions

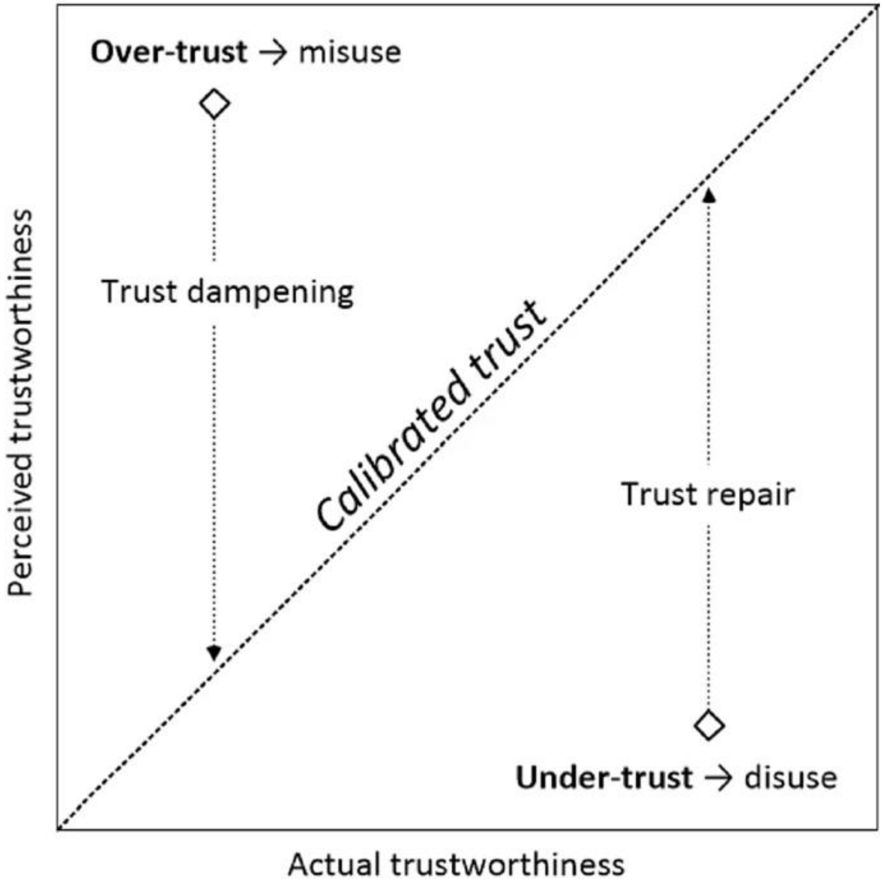
Trust: "…the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"*

Trustworthiness: how well one agent (e.g., the autonomy) *is perceived* to perform or *does perform* in a given situation (perceived vs actual trustworthiness)

Disuse: results from under-trust – i.e., not using the autonomy when one should

Misuse: results from over-trust – i.e., deferring to the autonomy when one shouldn't

Trust Calibration: the process of balancing user trust to and ideal level (minimize disuse and misuse)



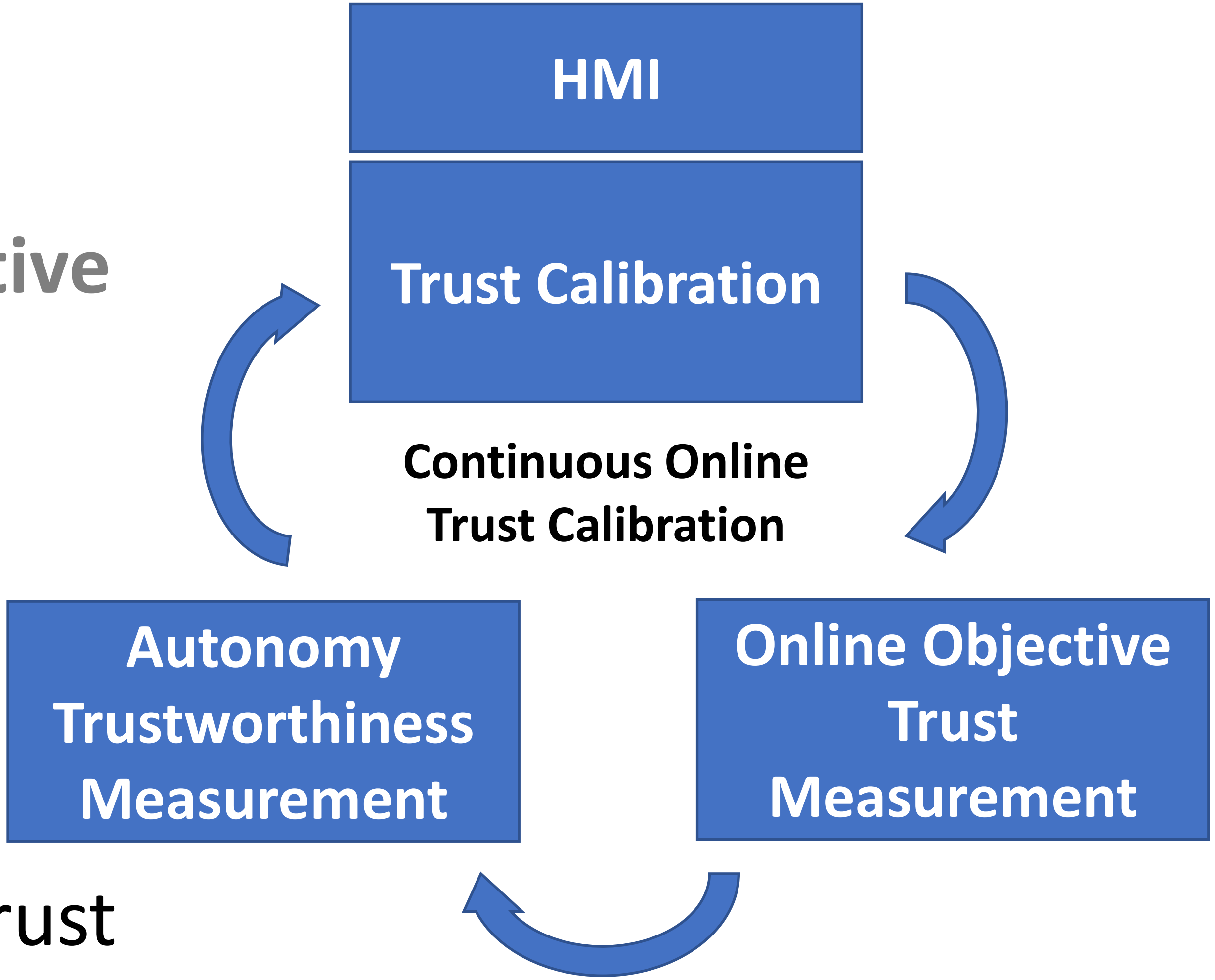
De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020).

*(Lee & See, 2004/Hoff & Bashir, 2015)

This information is approved for release

Online Trust Calibration

- Requires:
- Online measurement of **objective** human user **trust**
- Assessment of **autonomy trustworthiness**
- **Deliberately calibrating** user trust


HMI
Trust Calibration
Continuous Online Trust Calibration
Autonomy Trustworthiness Measurement
Online Objective Trust Measurement

This information is approved for release

Detecting Miscalibrated Trust

- Goal: detect over- or under- trust situations
- 3 parameters:

Pauto =
reliability of the agent
(probability that a task done by an agent will be successful)

Ptrust =
user's trust in the agent
(user's estimation of Pauto)

Pman =
capability of the user
(probability that a task done manually by a user will be successful)

- Over-trust = the user estimates that the agent is better at the task than the user, even though the actual reliability of the agent is lower than the user's capability

(Ptrust > Pman) & (Pman > Pauto)

- Under-trust = the user estimates that they are better at the task than the agent, even though the actual reliability of the agent is higher than the user's capability

(Ptrust < Pman) & (Pman < Pauto)

Okamura, K., & Yamada, S. (2020)

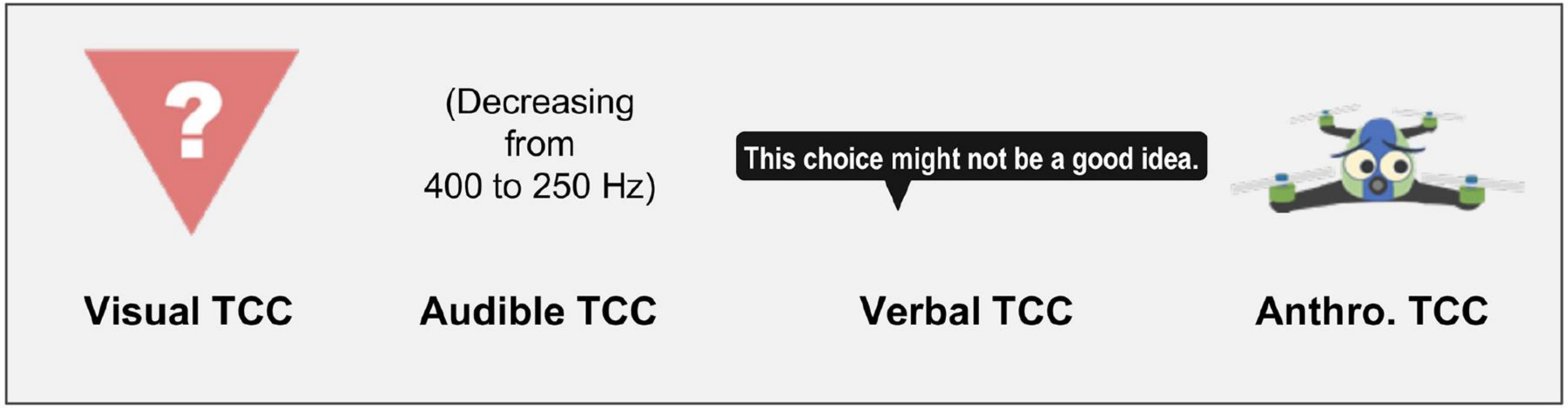
Responding to Mis-calibrated Trust: Two Approaches

- Two main approaches in literature:
  - Transparency / User Interface Adaptation
  - AI adaptation – change behavior of AI
    - E.g., Xu & Dudek (2016)

This information is approved for release

Trust Calibration Cues Study: Okamura & Yamada

- **Approach:** use one of 4 different TCCs to inform user about quality of autonomy in a task
- Guideline: TCCs should be noticeable in the task environment, should link the user to the next possible actions in the task
- Guideline: Only present TCCs when it is clear that trust is in need of calibration (rather than continuously)
- Result: Including any kind of TCC improved trust calibration over the group with no TCC

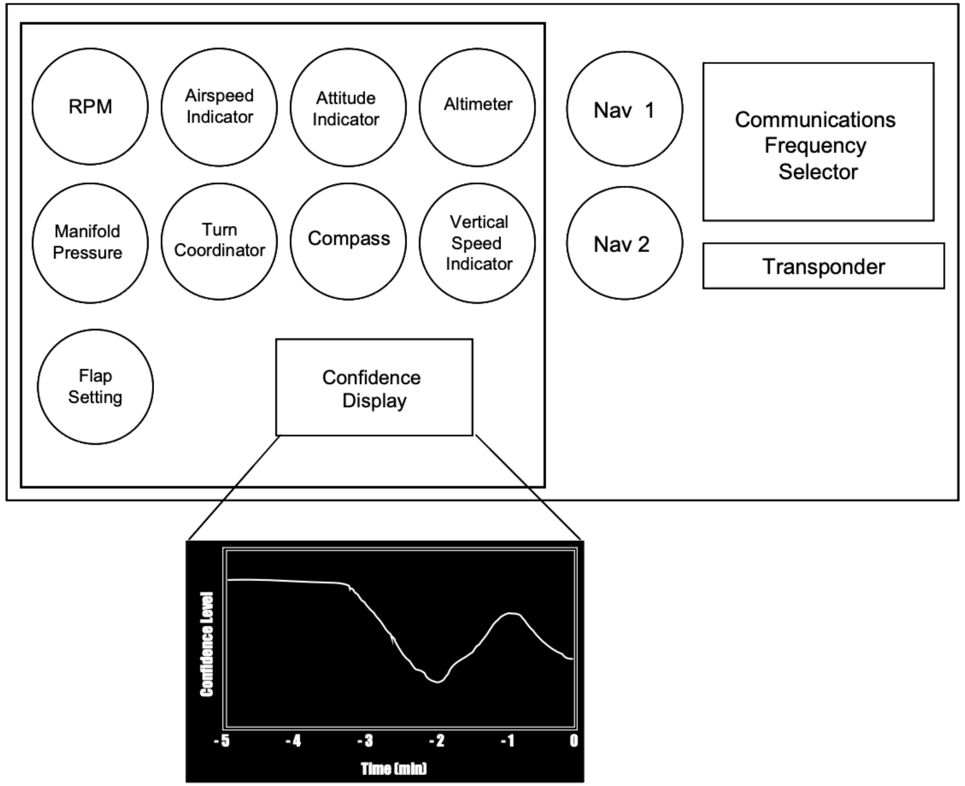

?
(Decreasing from 400 to 250 Hz)
This choice might not be a good idea.
Visual TCC
Audible TCC
Verbal TCC
Anthro. TCC

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *Plos one*, *15*(2), e0229132.

This information is approved for release

Dynamic System Confidence Display: McGuirl & Sarter

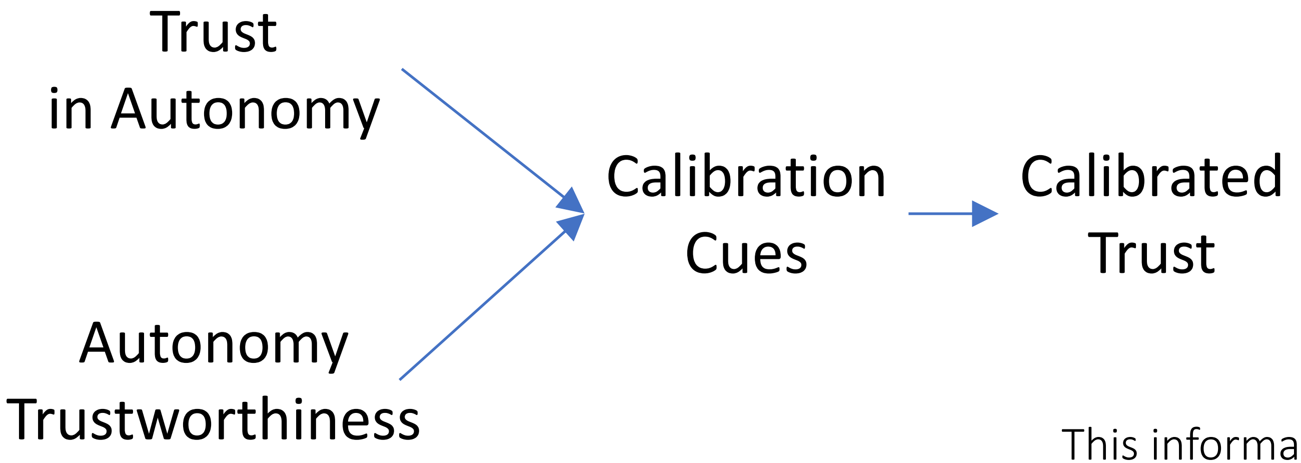
- **Approach:** present continually updated information about a system's confidence in its ability to perform assigned tasks
- "Confidence trend display": system's current confidence level and confidence over time
- Shown to improve trust calibration over systems that only present information about overall reliability



McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, *48*(4), 656-665.

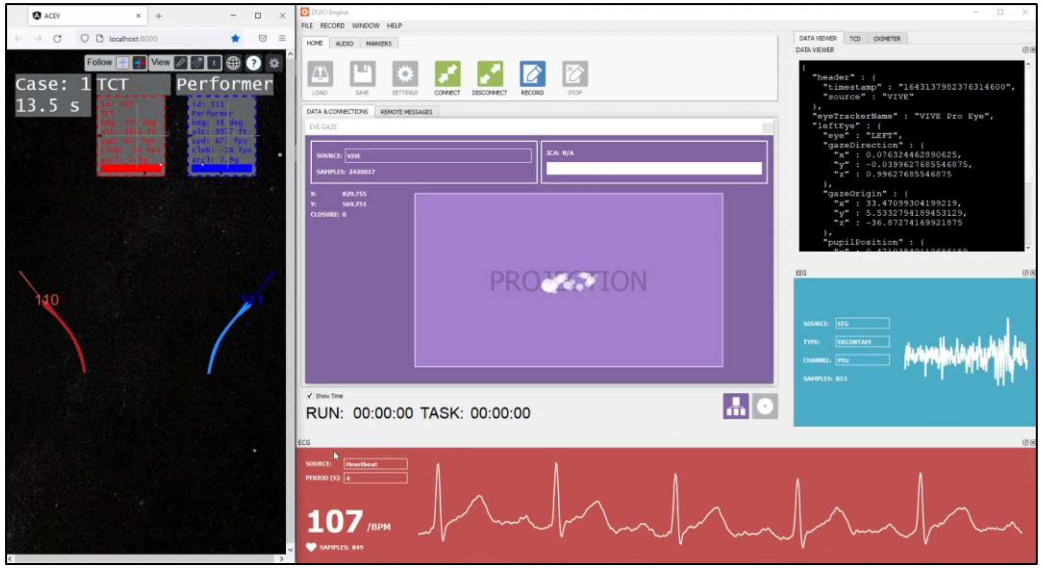
Context: Calibrating Pilot Trust in Dogfight Autonomy

- Aircraft is nominally controlled by autonomy during a dogfight
- Pilot can take over whenever desired but is also busy with other tasks
- Online measurement of trust:
  - Physiological sensors (e.g., heart rate, GSR, eye tracking, etc.)
  - Behavioral: taking over control, attention to other tasks
- Online trustworthiness assessment
  - 3rd party assessment based on prior performance in similar situations



This information is approved for release

TrustMATE^{TM} Testbed



Case: 1 TCT Performer
13.5 s

id: 111
Performer
hdg: 38 deg
alt: 9977 ft
spd: 673 fps
clmb: -18 fps
accl: 2.9g

SOURCE: VIVE
SAMPLES: 2420017
ICA: N/A
X: 829.755
Y: 569.751
CLOSURE: 0

✓ Show Time

RUN: 00:00:00 TASK: 00:00:00

```
{
  "header" : {
    "timestamp" : "1643137982376314600",
    "source" : "VIVE"
  },
  "eyeTrackerName" : "VIVE Pro Eye",
  "leftEye" : {
    "eye" : "LEFT",
    "gazeDirection" : {
      "x" : 0.076324462890625,
      "y" : -0.039962768554687\u0035,
      "z" : 0.99627685546875
    },
    "gazeOrigin" : {
      "x" : 33.47099304199219,
      "y" : 5.533279418945312\u0039,
      "z" : -36.87274169921875
    },
    "pupilPosition" : {
```

SOURCE: EEG
TYPE: DECONTAM
CHANNEL: POz
SAMPLES: 853

SOURCE: Heartbeat
PERIOD (S): 4
107 /BPM
SAMPLES: 849

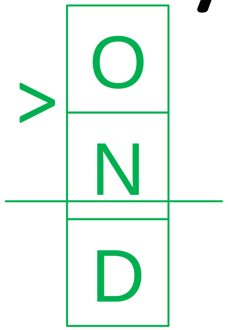
Purpose and Basis for HMI Designs

- Purpose: To inform a pilot about the autonomy to help facilitate (and calibrate) trust in the autonomy

- Basis:
  - Pilot interviews
    - Safety – Performance – Situation Assessment … and trends
  - Literature on trust
    - McGuirl & Sarter (2006); Okamura, K., & Yamada, S. (2020)
  - Existing cockpit (HMD) displays
    - Existing and proposed fighter cockpit displays / current practices

Trust Calibration Cues via Transparency

- **Situation**: Offensive/Neutral/Defensive (O/N/D) indicator (with trend) as determined by autonomy



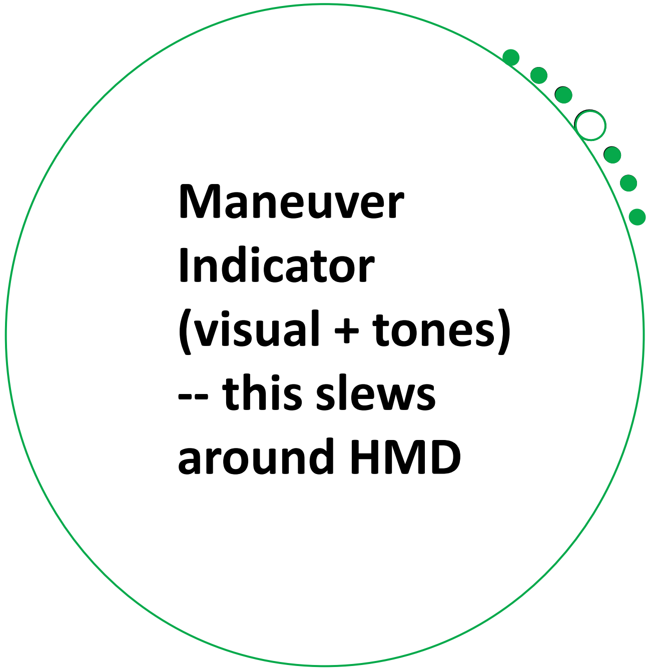
Situation Indicator

- **Performance**: Win estimation (trustworthiness) as determined by 3rd party assessor

W30: 75

Win Estimation
75% chance of win in next 30 seconds

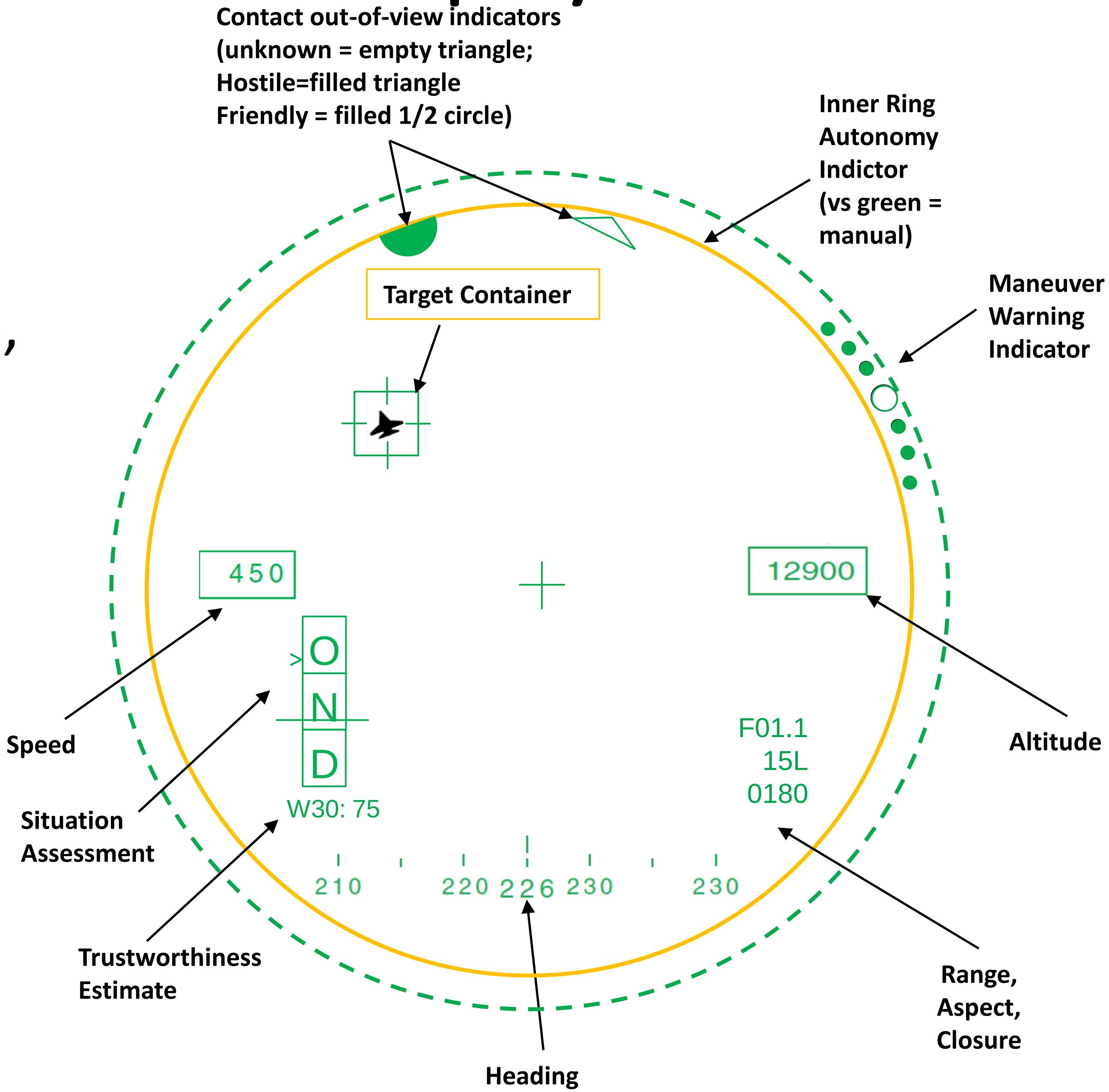
- **Safety**: G-maneuver indicator
  - Indicates direction and intensity of movement (e.g., a turn) some number of seconds in the future
  - Voice indicator also when predicted Gs exceeded threshold



Maneuver Indicator (visual + tones) -- this slews around HMD

Info Delivery: Head-Mounted Display

- Built in virtual reality (Unity3D portrayed in HTC Vive Pro Eye)
- Trust cues overlaid on top of other flight info – e.g., altitude, heading, etc.
- Autonomy mode indicator (orange vs green)
- Always in front of user even when turning head
- Not a lot of real estate to add visual cues



Contact out-of-view indicators (unknown = empty triangle; Hostile=filled triangle Friendly = filled 1/2 circle)

Inner Ring Autonomy Indictor (vs green = manual)

Maneuver Warning Indicator

Target Container

Speed

Situation Assessment

Trustworthiness Estimate

Heading

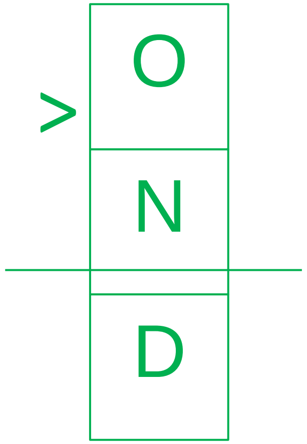
Range, Aspect, Closure

Altitude

This information is approved for release

Early Findings

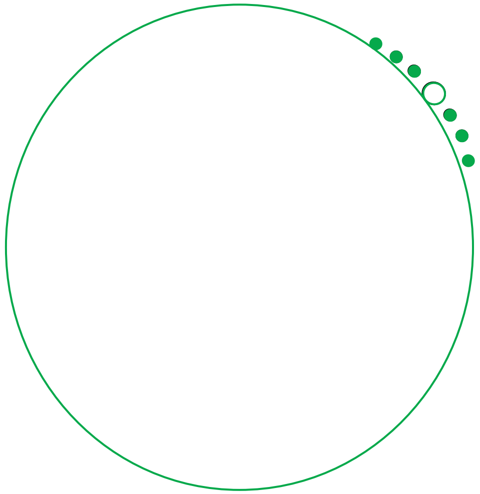
- **Situation Indicator** was deemed useful by pilots
  - Some requests to only show it when defensive (matches Okamura & Yamata)



- **Win estimation** challenging to compute, needs to be reliable to be useful; probably needs to be some combination of self-assessment and 3rd party

W30: 75

- **Maneuver warning:** Not fully functional at evaluation
  - Anticipated that it would be more useful in real flight where real Gs happen



Lessons Learned

- High dependence on autonomy providers to give useful, reliable info
  - Trustworthiness ratings, assessment of situation, lookahead predictions
  - Not all autonomy implementations produce equivalent data

- The rate of info change must be dampened to user perception speeds
  - Autonomy in constant reappraisal, many results in sub-second timeframe

Summary

- Online trust calibration requires:
  - Continuously measuring human trust in autonomy
  - Continuously measuring trustworthiness of autonomy
  - Continuously computing current level of trust calibration and manipulating the HMI

- Lots of different ways to manipulate HMI that could impact trust
  - The HMI itself could negatively affect trust even if the autonomy performs well
  - Counterintuitive, but to calibrate, must sometimes tell user to trust the system less

Thanks!

SOARTECH

Modeling human reasoning.
Enhancing human performance.