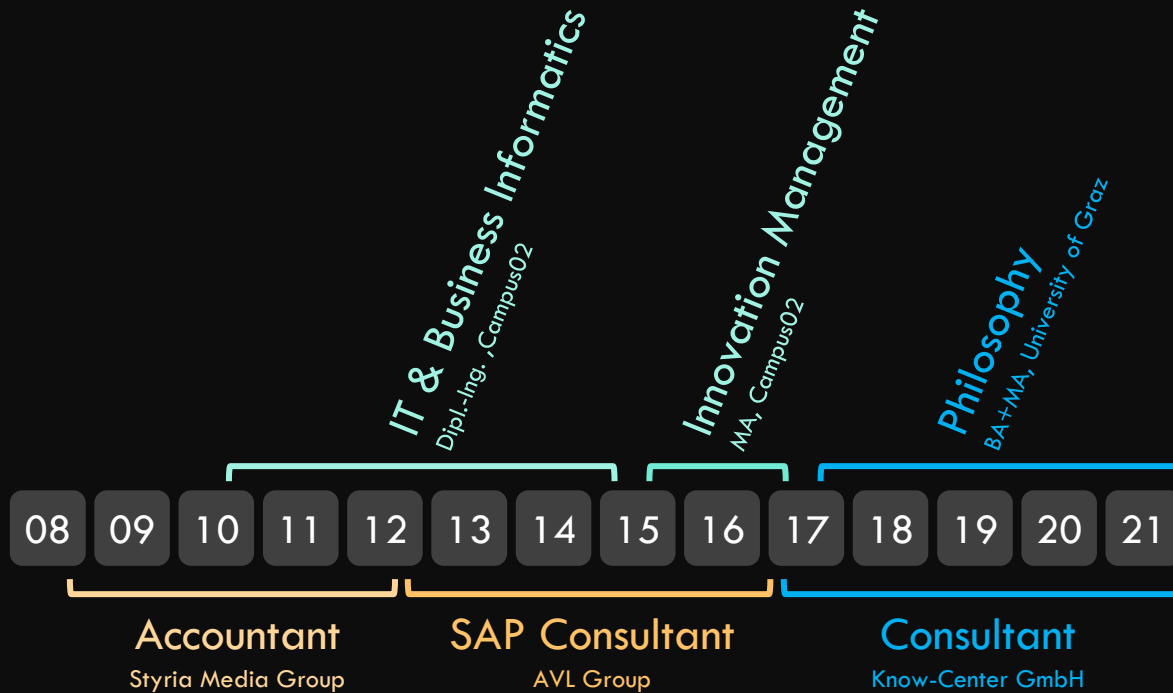


MAKE US SMILE!

AI AND THE VIOLATION OF HUMAN INTENTIONS



Christof Wolf-Brenner

christof.brenner@gmx.at
cbrenner@know-center.at

OUTLINE

Motivation

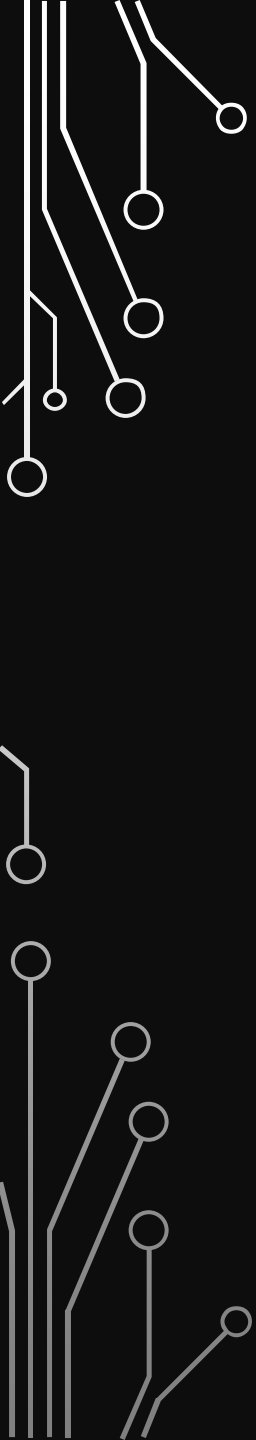


Argument



Impact & Outlook

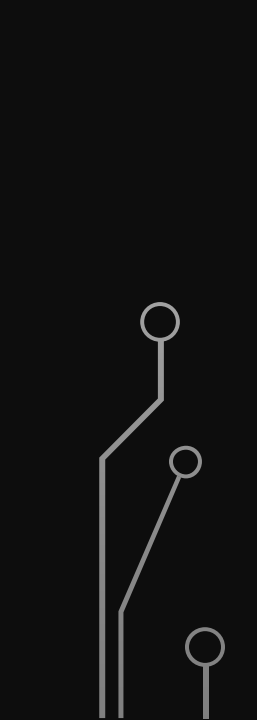
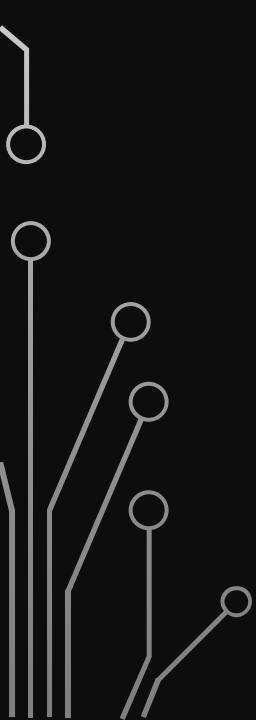
1997 08 04 02:14 AM







PERVERSE INSTANTIATION



An unintended way through which AI figures out to achieve some goal



unintended way

paralyze facial muscles

extermination

breed cobras

run away endlessly

make ham sandwich for vegetarian



some goal

make us smile

defend against threats

reduce cobra population

win at hide & seek

make a sandwich



The image features a dark background with white and teal decorative elements. In the corners, there are stylized circuit board traces with circular nodes. The main text is centered and consists of two lines: the first line is in white and the second line is in teal.

UNINTENDED WAYS TO ACHIEVE GOALS ARE NOT ALWAYS BAD

BUT WE NEED TO FIGURE OUT HOW TO
DISTINGUISH BENIGN FROM MALIGN CASES

THE ORIGIN OF UNINTENDED

Goals

- Underspecified
- Misinterpreted



THE ORIGIN OF UNINTENDED

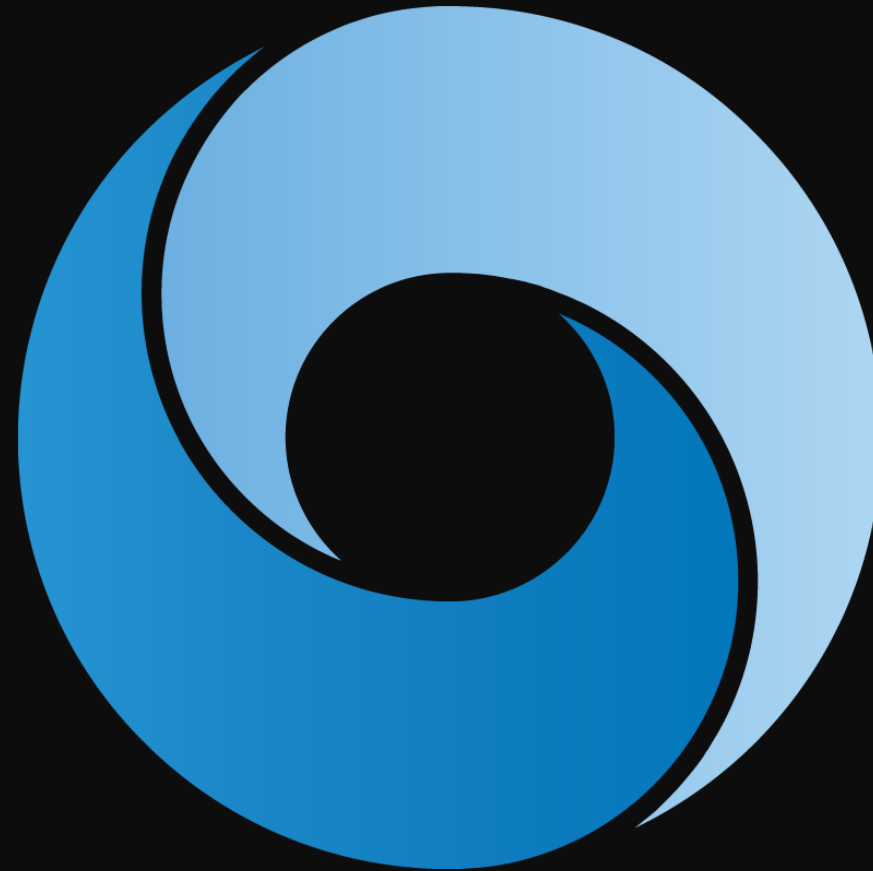


Ways to achieve goals

- Unintended actions or combinations of actions
 - Unintended interaction with environment
- 
- 

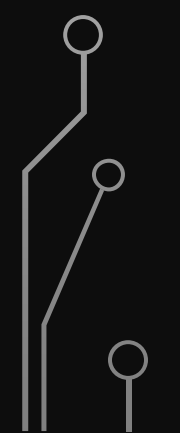
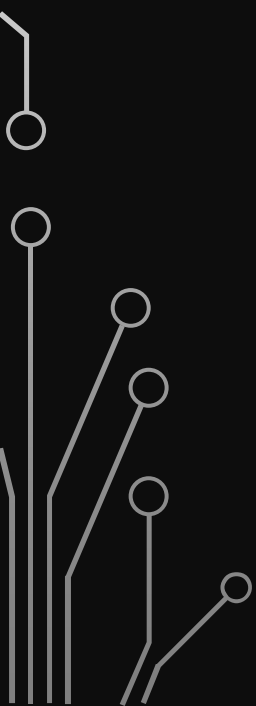
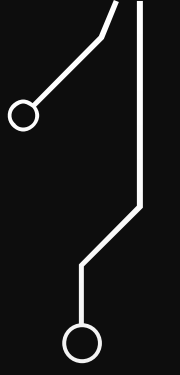
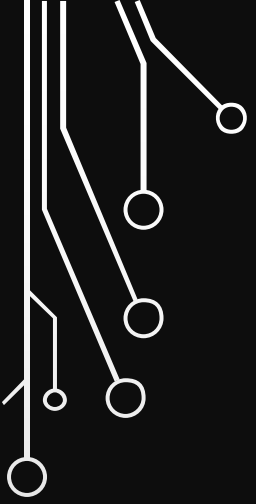
The image features a black background with white, stylized circuit board traces in the four corners. These traces consist of straight lines of varying lengths and angles, ending in small white circles that represent components or connection points. The patterns are symmetrical and decorative, framing the central text.

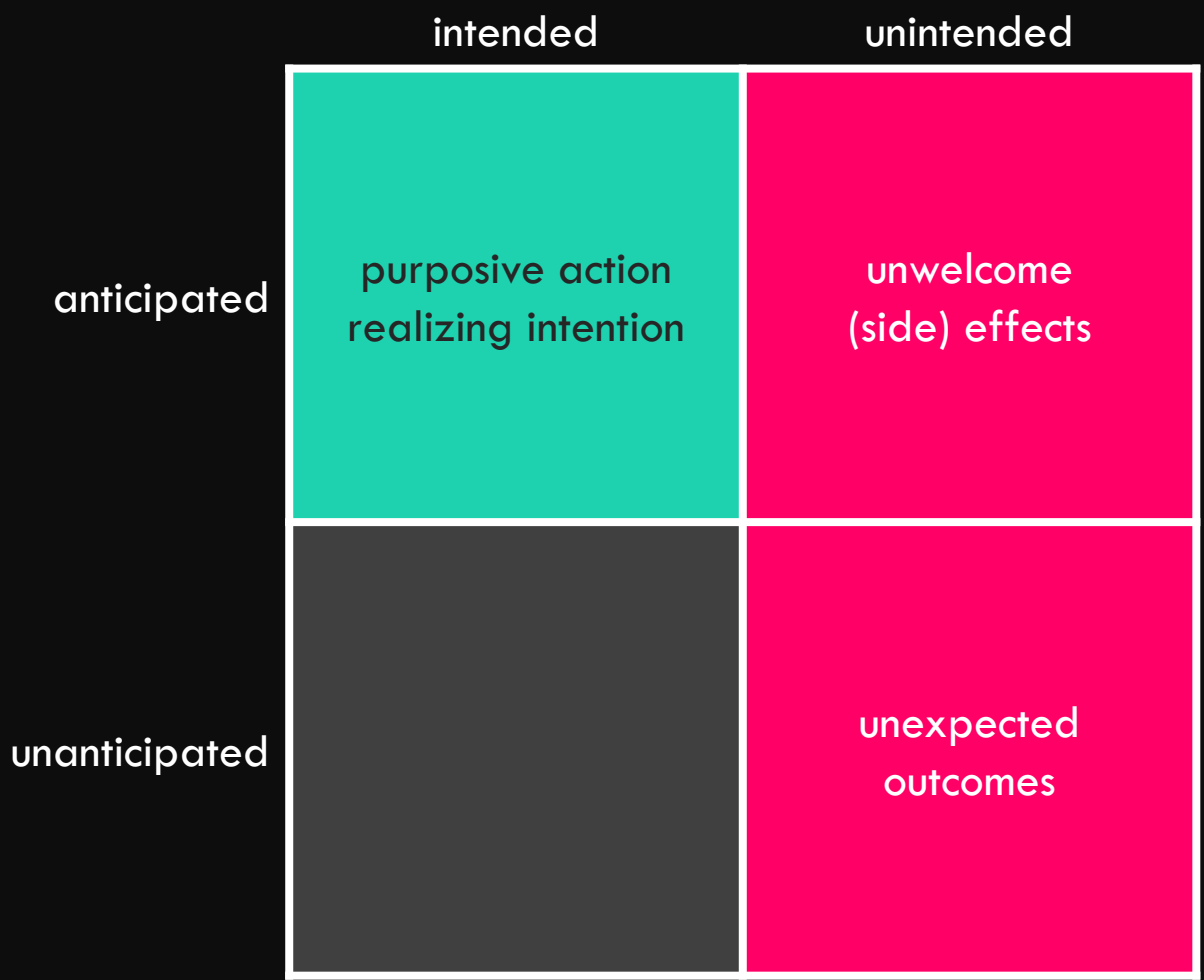
(NOT) INTENDED



Silver et al (2021) – Reward Is Enough

<https://doi.org/10.1016/j.artint.2021.103535>





Zwart (2015) – Unintended but not Unanticipated Consequences

<http://dx.doi.org/10.1007/s11186-015-9247-6>

*THANK YOU
FOR YOUR ATTENTION*



Christof Wolf-Brenner

christof.brenner@gmx.at

cbrenner@know-center.at