WE ARE AI
#3

Who lives, Who dies, Who decides?

TERMS OF USE

All the panels in this comic book are licensed CC BY-NC-ND 4.0. Please refer to the license page for details on how you can use this artwork.

TL;DR: Feel free to use panels/groups of panels in your presentations/articles, as long as you

1. Provide the proper citation
2. Do not make modifications to the individual panels themselves

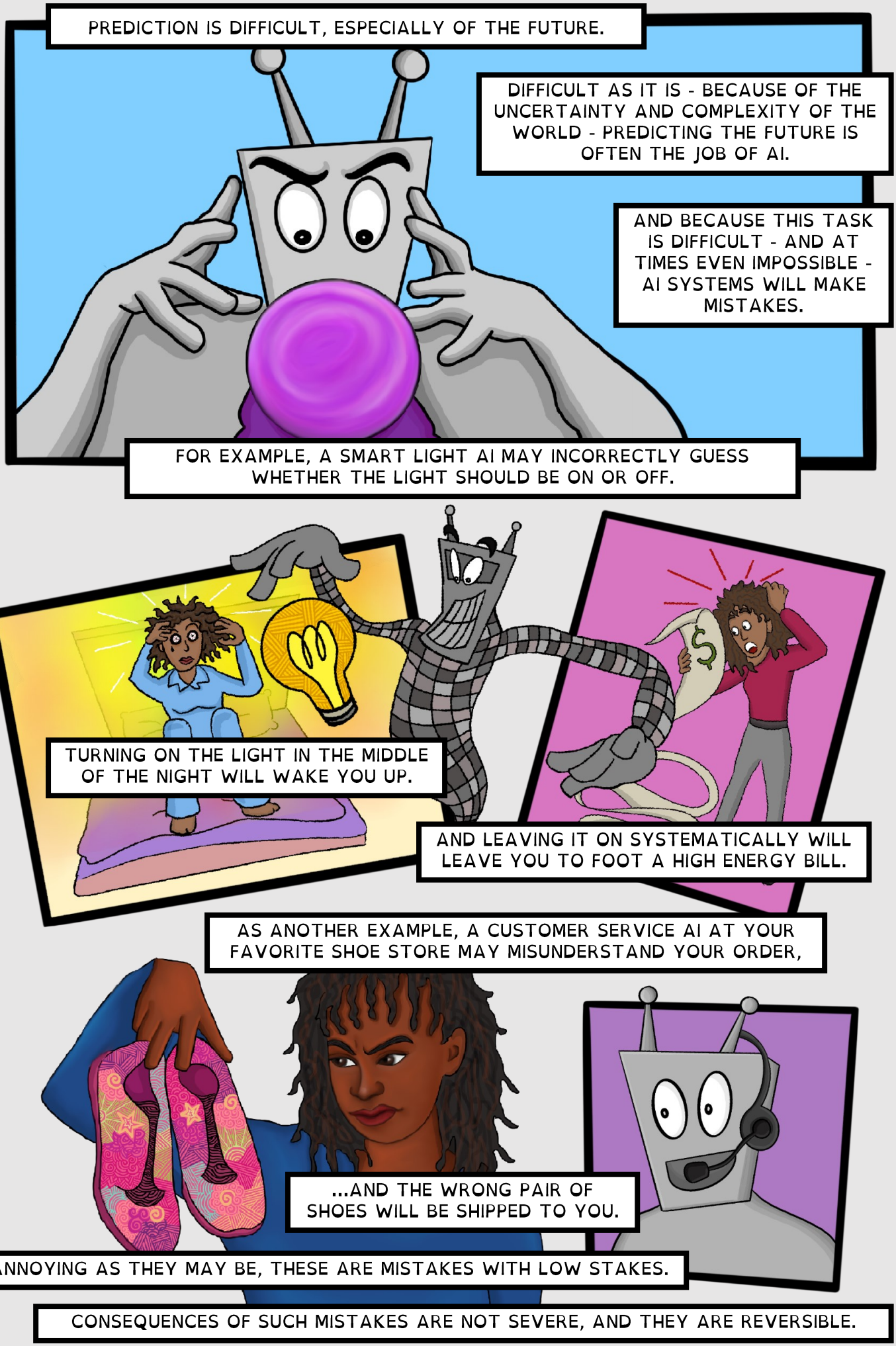
Cite as:

Julia Stoyanovich, Mona Sloane and Falaah Arif Khan.
"Who lives, who dies, who decides?". *We are AI Comics*, Vol 3 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol3_en.pdf

Contact:

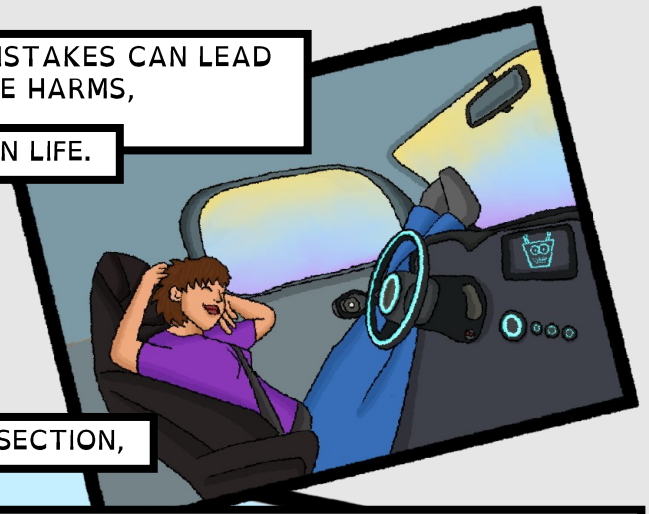
Please direct any queries about using elements from this comic to themachinelearnist@gmail.com and cc stoyanovich@nyu.edu

Licensed CC BY-NC-ND 4.0


PREDICTION IS DIFFICULT, ESPECIALLY OF THE FUTURE.
DIFFICULT AS IT IS - BECAUSE OF THE UNCERTAINTY AND COMPLEXITY OF THE WORLD - PREDICTING THE FUTURE IS OFTEN THE JOB OF AI.
AND BECAUSE THIS TASK IS DIFFICULT - AND AT TIMES EVEN IMPOSSIBLE - AI SYSTEMS WILL MAKE MISTAKES.
FOR EXAMPLE, A SMART LIGHT AI MAY INCORRECTLY GUESS WHETHER THE LIGHT SHOULD BE ON OR OFF.
TURNING ON THE LIGHT IN THE MIDDLE OF THE NIGHT WILL WAKE YOU UP.
AND LEAVING IT ON SYSTEMATICALLY WILL LEAVE YOU TO FOOT A HIGH ENERGY BILL.
AS ANOTHER EXAMPLE, A CUSTOMER SERVICE AI AT YOUR FAVORITE SHOE STORE MAY MISUNDERSTAND YOUR ORDER,
...AND THE WRONG PAIR OF SHOES WILL BE SHIPPED TO YOU.
ANNOYING AS THEY MAY BE, THESE ARE MISTAKES WITH LOW STAKES.
CONSEQUENCES OF SUCH MISTAKES ARE NOT SEVERE, AND THEY ARE REVERSIBLE.

HOWEVER, THERE ARE CASES WHERE MISTAKES CAN LEAD TO CATASTROPHIC IRREVERSIBLE HARMS,

...EVEN TO THE LOSS OF HUMAN LIFE.



CONSIDER AN AUTONOMOUS CAR -

AN AI THAT IS ABOUT TO CROSS AN INTERSECTION,

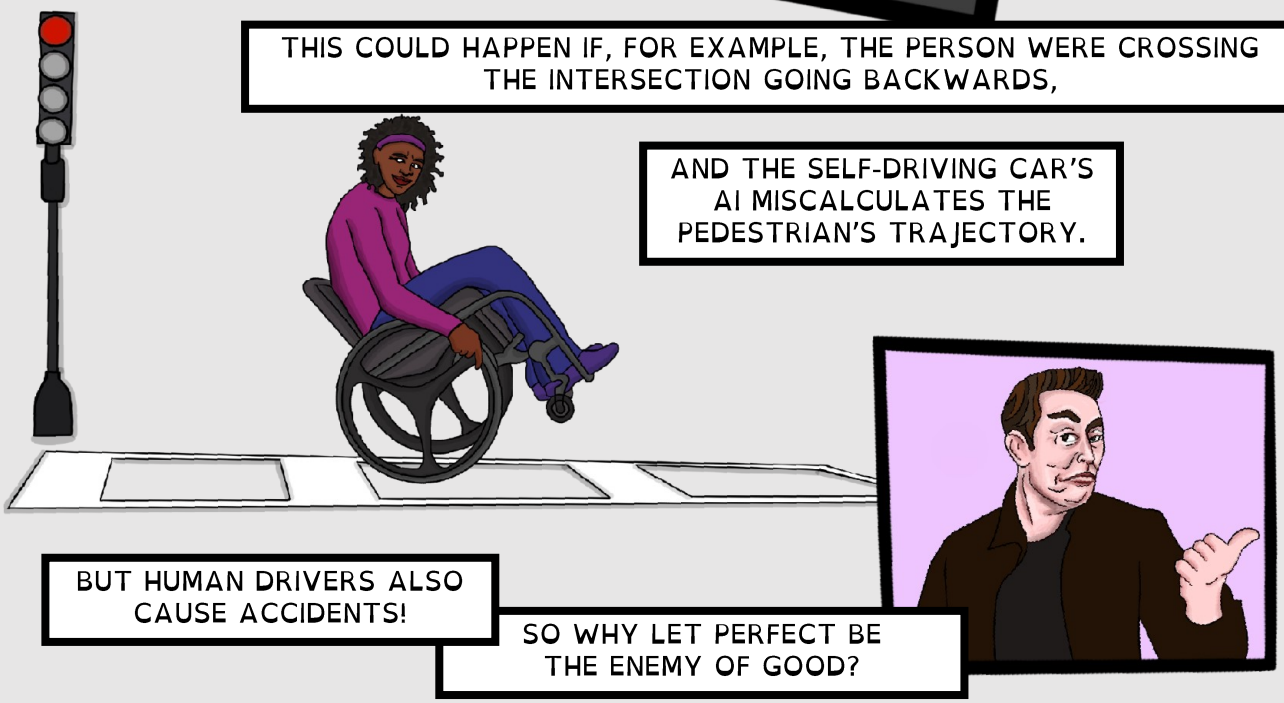
AND THAT DOES NOT RECOGNIZE A PERSON ON A BICYCLE AS ONE OF THE TYPES OF OBJECTS IT WOULD EXPECT TO SEE ON THE ROAD.

THE CAR WOULD THEN CONTINUE ON ITS PATH, RUNNING THE CYCLIST OVER.

ANOTHER EXAMPLE IS WHEN THE AUTONOMOUS CAR DOES NOT DETECT THE PRESENCE OF A PERSON IN A WHEELCHAIR CROSSING THE INTERSECTION.

THIS COULD HAPPEN IF, FOR EXAMPLE, THE PERSON WERE CROSSING THE INTERSECTION GOING BACKWARDS,

AND THE SELF-DRIVING CAR'S AI MISCALCULATES THE PEDESTRIAN'S TRAJECTORY.



BUT HUMAN DRIVERS ALSO CAUSE ACCIDENTS!

SO WHY LET PERFECT BE THE ENEMY OF GOOD?

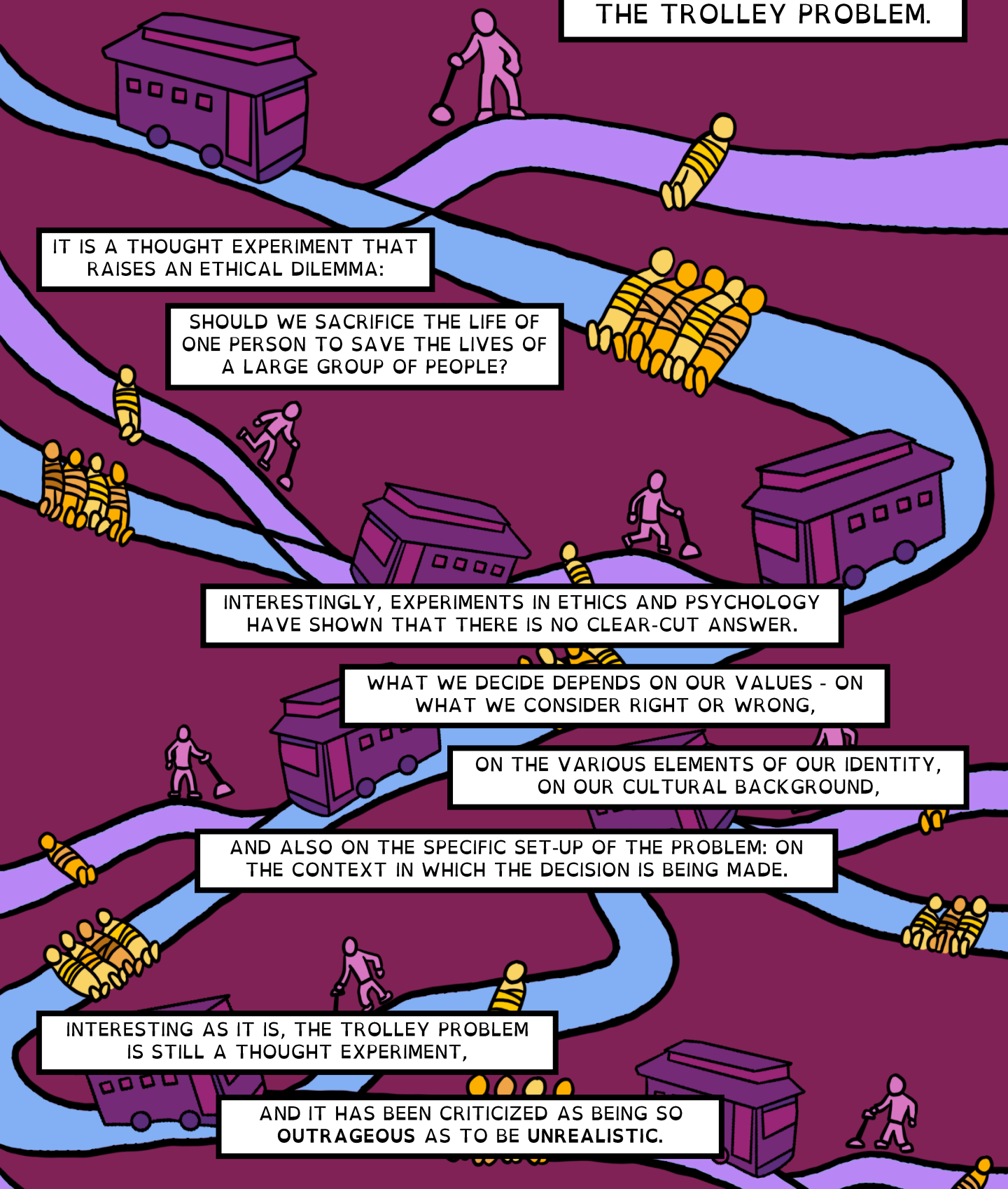
SHOULDN'T WE BE PREPARED TO SUFFER A FEW MISTAKES MADE BY AUTONOMOUS CARS IN THE NAME OF INCREASED OVERALL SAFETY OF OUR TRANSPORTATION SYSTEM, AND THE CONVENIENCE TO THE DRIVERS?

IN FACT, CAN'T WE ENCODE OUR JUDGEMENT ABOUT WHAT MISTAKES ARE MORE IMPORTANT TO AVOID, AND LET AN AI SORT OUT THE TRADE-OFFS?

CAN'T WE EQUIP OUR AI WITH VALUES?

A FAMOUS EXAMPLE THAT MAKES US THINK ABOUT OUR VALUES, AND TRADE-OFFS THEY INTRODUCE, IS

THE TROLLEY PROBLEM.



IT IS A THOUGHT EXPERIMENT THAT RAISES AN ETHICAL DILEMMA:

SHOULD WE SACRIFICE THE LIFE OF ONE PERSON TO SAVE THE LIVES OF A LARGE GROUP OF PEOPLE?

INTERESTINGLY, EXPERIMENTS IN ETHICS AND PSYCHOLOGY HAVE SHOWN THAT THERE IS NO CLEAR-CUT ANSWER.

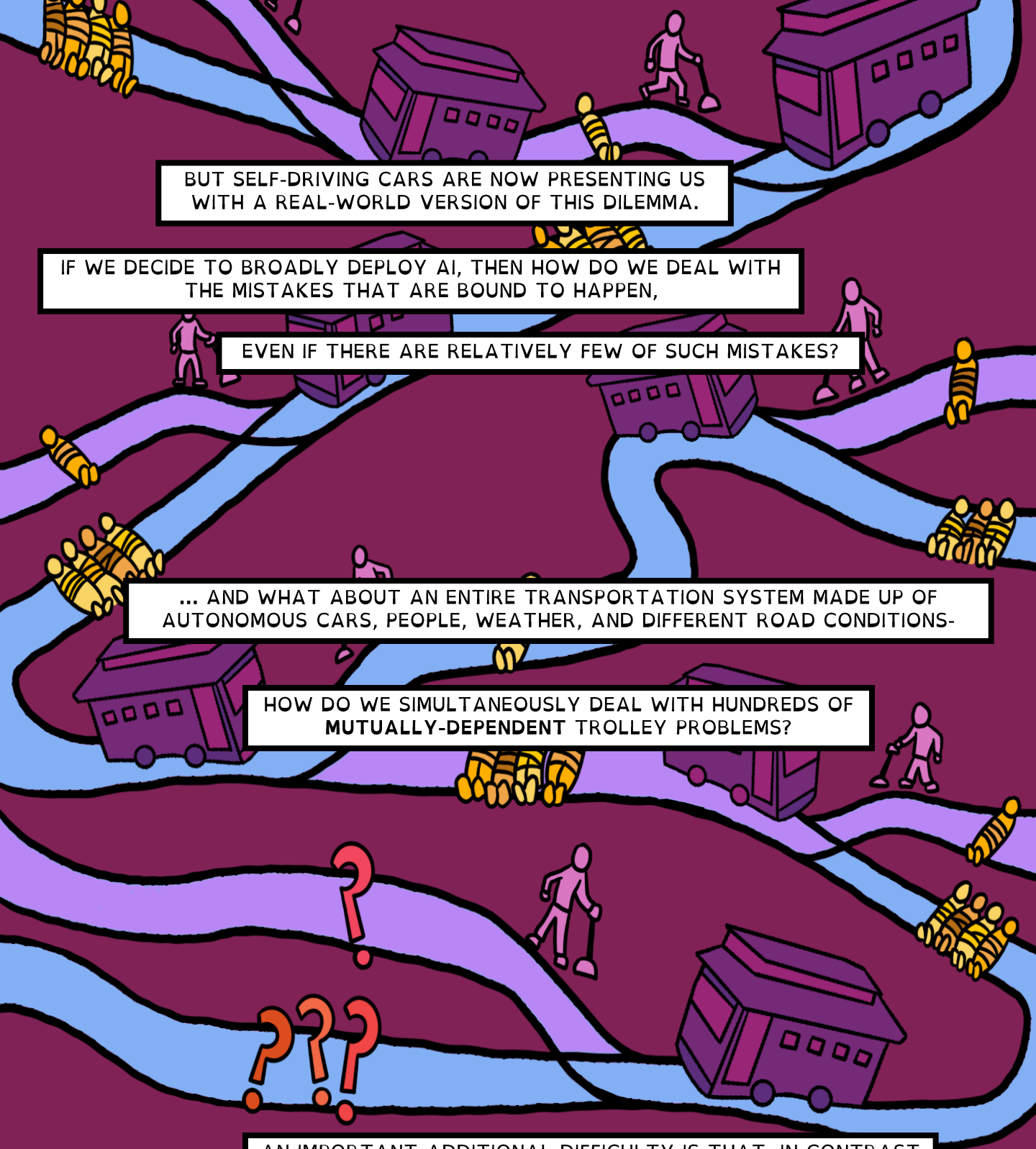
WHAT WE DECIDE DEPENDS ON OUR VALUES - ON WHAT WE CONSIDER RIGHT OR WRONG,

ON THE VARIOUS ELEMENTS OF OUR IDENTITY, ON OUR CULTURAL BACKGROUND,

AND ALSO ON THE SPECIFIC SET-UP OF THE PROBLEM: ON THE CONTEXT IN WHICH THE DECISION IS BEING MADE.

INTERESTING AS IT IS, THE TROLLEY PROBLEM IS STILL A THOUGHT EXPERIMENT,

AND IT HAS BEEN CRITICIZED AS BEING SO **OUTRAGEOUS** AS TO BE **UNREALISTIC.**



AN IMPORTANT ADDITIONAL DIFFICULTY IS THAT, IN CONTRAST TO THE CLASSIC TROLLEY PROBLEM, WHERE IT IS KNOWN HOW MANY PEOPLE ARE ON WHAT SIDE OF THE TRACK,

AN AUTONOMOUS CAR — AND OTHER TYPES OF TECHNOLOGY — OPERATE UNDER A HIGH DEGREE OF **UNCERTAINTY**.

IT MAY BE UNKNOWN WHETHER THERE ARE EVEN PEOPLE ON THE TRACKS,

LET ALONE HOW MANY OF THEM THERE ARE, AND WHICH GROUPS THEY MAY REPRESENT.

HOW DO WE MAKE VALUE JUDGMENTS IN THE FACE OF UNCERTAINTY?

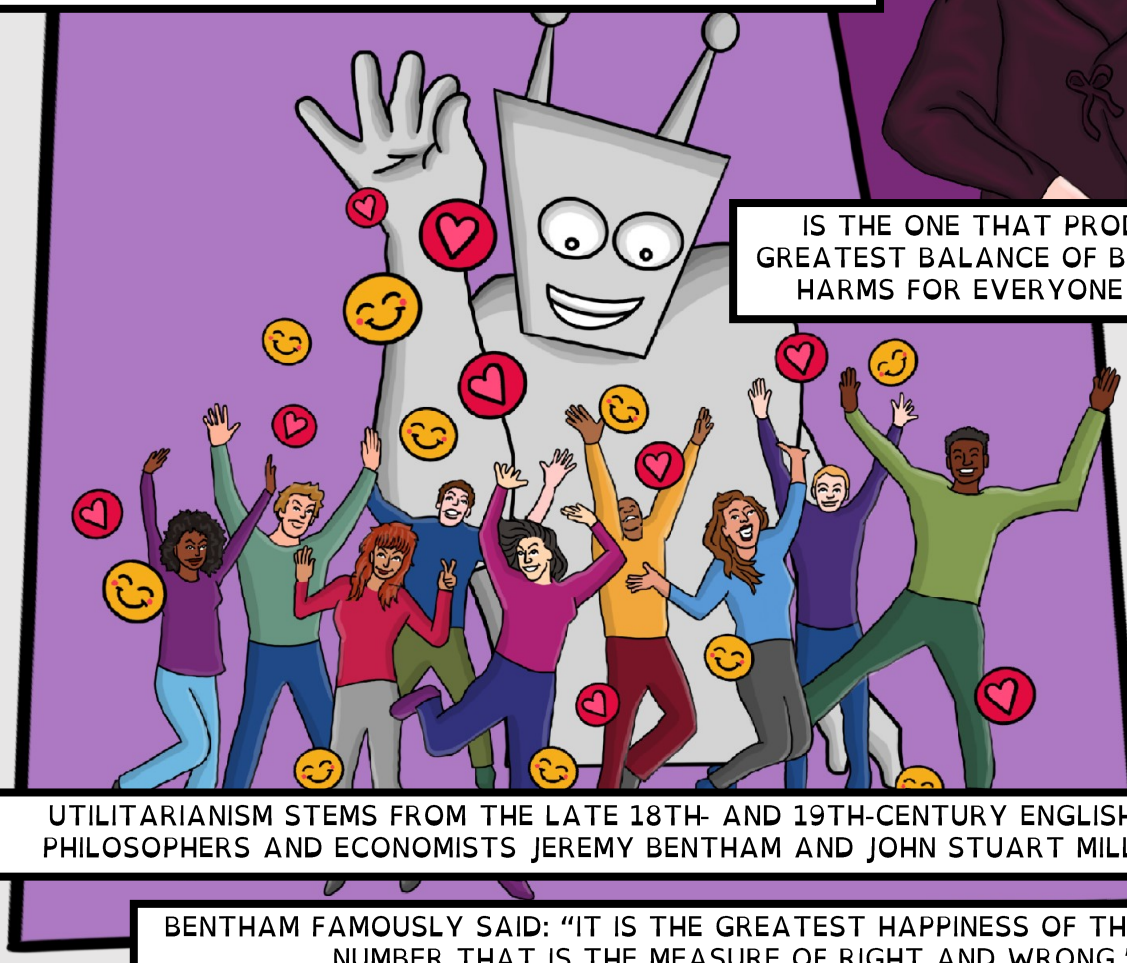
THE TROLLEY CAR PROBLEM ILLUSTRATES A SPECIFIC DOCTRINE OF MORAL PHILOSOPHY -

UTILITARIANISM.

PERHAPS THIS DOCTRINE CAN OFFER US SOME GUIDANCE?

UTILITARIANISM IS A MORAL PRINCIPLE THAT HOLDS THAT THE RIGHT COURSE OF ACTION — IN ANY SITUATION —

IS THE ONE THAT PRODUCES THE GREATEST BALANCE OF BENEFITS OVER HARMS FOR EVERYONE AFFECTED.

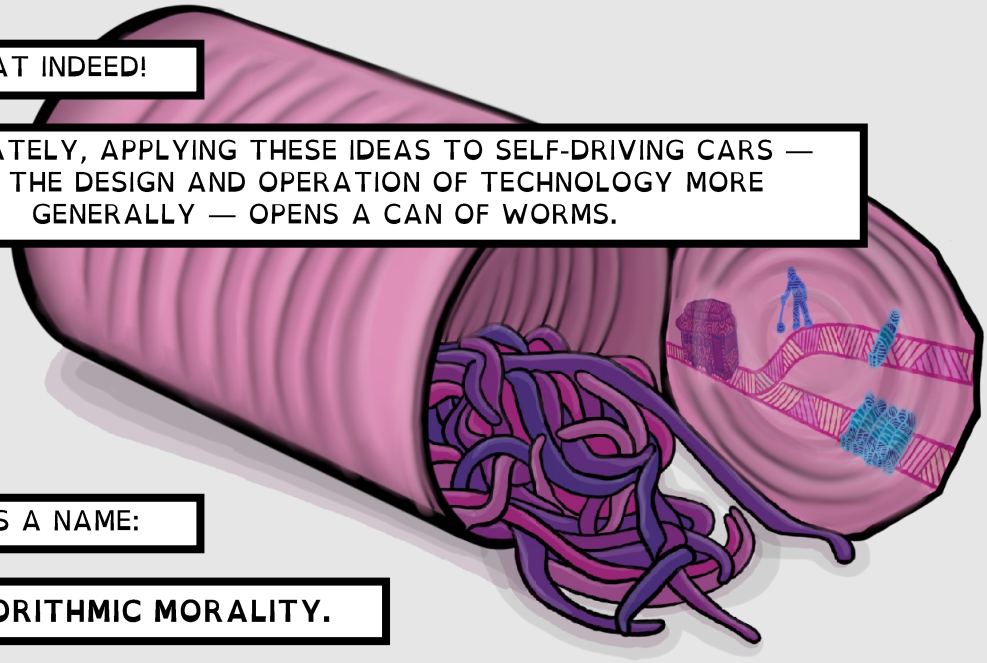


UTILITARIANISM STEMS FROM THE LATE 18TH- AND 19TH-CENTURY ENGLISH PHILOSOPHERS AND ECONOMISTS JEREMY BENTHAM AND JOHN STUART MILL.

BENTHAM FAMOUSLY SAID: "IT IS THE GREATEST HAPPINESS OF THE GREATEST NUMBER THAT IS THE MEASURE OF RIGHT AND WRONG."

SOUNDS GREAT INDEED!

UNFORTUNATELY, APPLYING THESE IDEAS TO SELF-DRIVING CARS — AND TO THE DESIGN AND OPERATION OF TECHNOLOGY MORE GENERALLY — OPENS A CAN OF WORMS.



AND IT HAS A NAME:

ALGORITHMIC MORALITY.

ALGORITHMIC MORALITY IS THE ACT OF ATTRIBUTING MORAL REASONING TO ALGORITHMS.

DOING SO IS PROBLEMATIC. HERE IS WHY.

TO START, HOW DO WE MEASURE HAPPINESS AND UNHAPPINESS?

AND HOW DO WE THEN ENCODE THESE MEASUREMENTS INTO A SET OF OBJECTIVES THAT AN ALGORITHM WILL UNDERSTAND?

THERE RARELY EXISTS A MATHEMATICAL FORMULA OR A LOGICAL STATEMENT THAT CAN CAPTURE THE BALANCE BETWEEN THE BENEFITS AND THE HARMS.

IN OTHER WORDS: THERE SIMPLY ISN'T A FORMULA FOR "RIGHT" OR "WRONG".

AND THERE ISN'T A FORMULA FOR VALUES, AND FOR HOW VALUES EMERGE AND CHANGE IN COMPLEX SOCIAL SITUATIONS.

ANOTHER REASON WHY ALGORITHMIC MORALITY IS PROBLEMATIC IS THAT,

WHEN A MISTAKE IN JUDGMENT ABOUT WHAT IS RIGHT OR WRONG IS MADE,

— AND, AS WE ALREADY KNOW, MISTAKES WILL BE MADE BECAUSE THE WORLD IS COMPLEX, UNCERTAIN, AND PERHAPS EVEN UNPREDICTABLE —

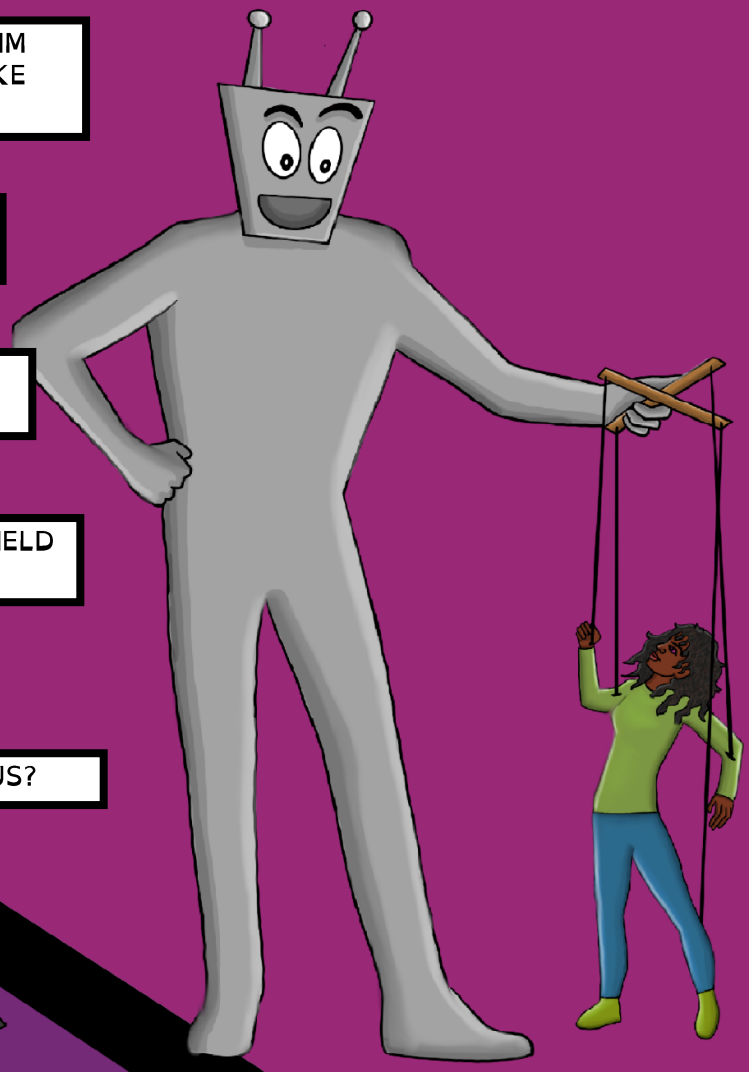
ALGORITHMIC MORALITY WOULD REQUIRE AN ALGORITHM TO TAKE RESPONSIBILITY FOR THE MISTAKE.

BUT HOLDING AN ALGORITHM RESPONSIBLE FOR A MISTAKE MAKES NO SENSE:

AN ALGORITHM DOES NOT POSSESS CONSCIOUSNESS OR FREE WILL,

IT DOES NOT MAKE AN INTENTIONAL CHOICE THAT LEADS TO A MISTAKE,

AND SO CANNOT BE HELD ACCOUNTABLE.



WHERE DOES THIS LEAVE US?

THE CAN-OPENER THAT IS THE TROLLEY PROBLEM SHOWED US THAT WE CANNOT DELEGATE ETHICS TO MACHINES.

THAT IT IS STILL UP TO US, HUMANS, TO MAKE CHOICES AND TAKE ACTIONS (OR CHOOSE NOT TO ACT),

IN ACCORDANCE WITH OUR VALUES, AND WITH EXISTING LAWS.

AND THEN IT'S UP TO US TO TAKE RESPONSIBILITY FOR THE CONSEQUENCES OF ANY MISTAKES.

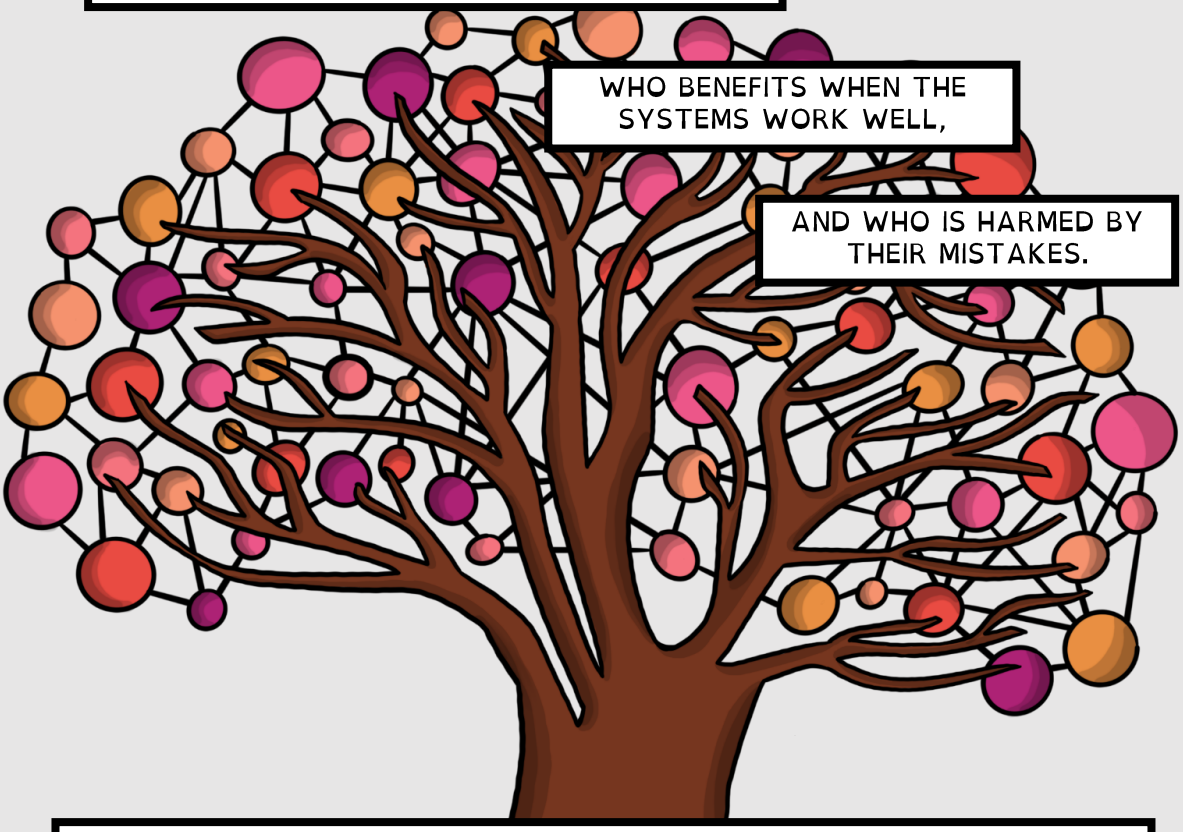
WE CANNOT OUTSOURCE THE WORK OF BEING HUMAN TO A MACHINE.

IN SUMMARY, TO EMBED ETHICS INTO SOCIO-TECHNICAL SYSTEMS SUCH AS AI,

WE MUST THINK ABOUT WHAT VALUES ARE BAKED INTO THESE SYSTEMS,

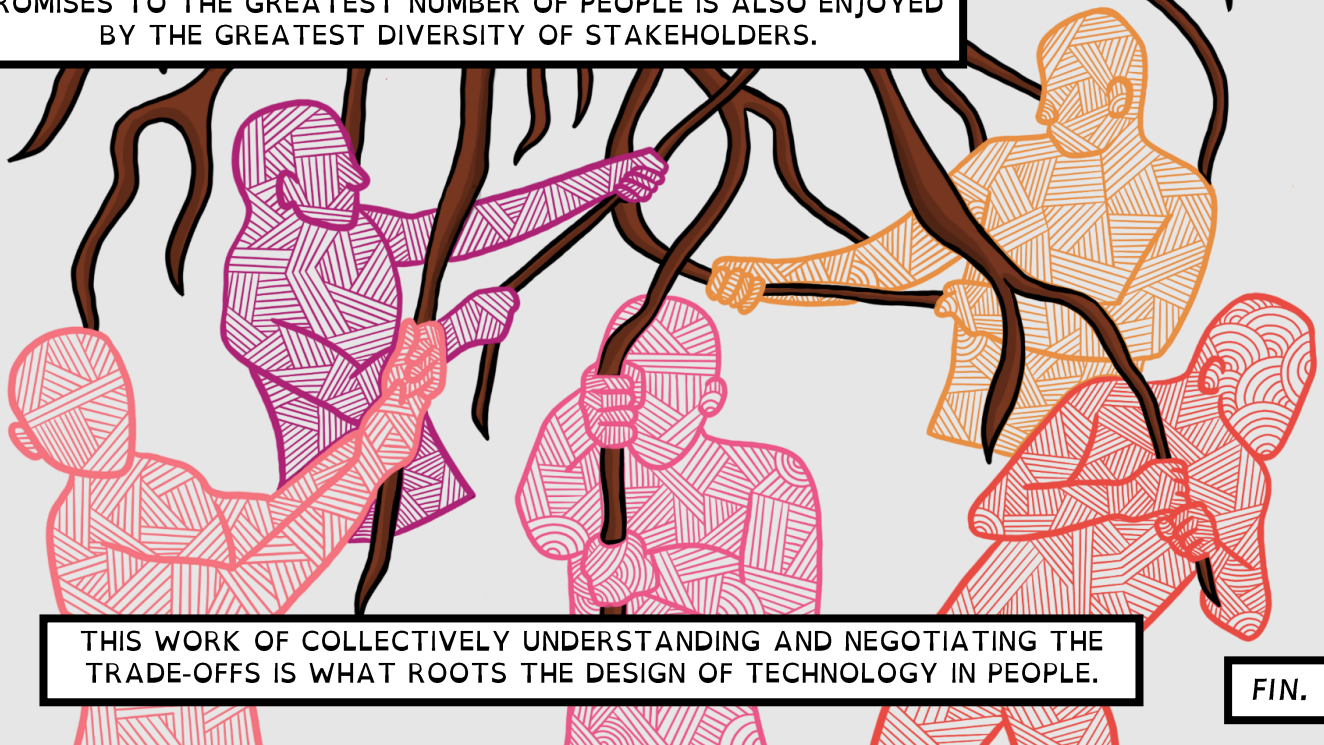
WHO BENEFITS WHEN THE SYSTEMS WORK WELL,

AND WHO IS HARMED BY THEIR MISTAKES.



AND WE MUST COLLECTIVELY TAKE RESPONSIBILITY FOR DECIDING ON THE BALANCE BETWEEN THE BENEFITS AND THE HARMS,

SO THAT "THE GREATEST HAPPINESS" THAT JEREMY BENTHAM PROMISES TO THE GREATEST NUMBER OF PEOPLE IS ALSO ENJOYED BY THE GREATEST DIVERSITY OF STAKEHOLDERS.



THIS WORK OF COLLECTIVELY UNDERSTANDING AND NEGOTIATING THE TRADE-OFFS IS WHAT ROOTS THE DESIGN OF TECHNOLOGY IN PEOPLE.

FIN.