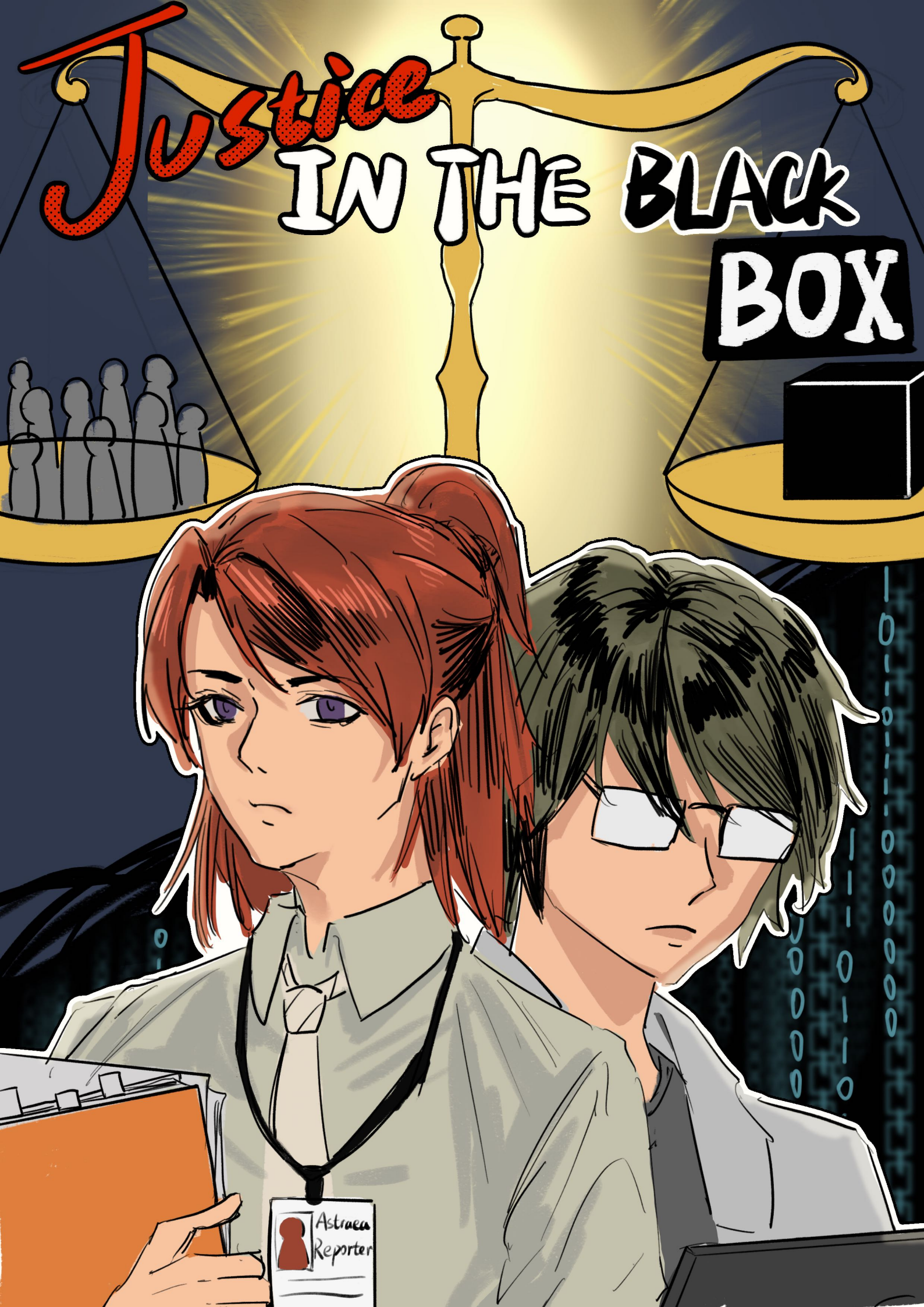

Justice
IN THE BLACK
BOX
Astraea
Reporter

Title: Justice in the Black Box

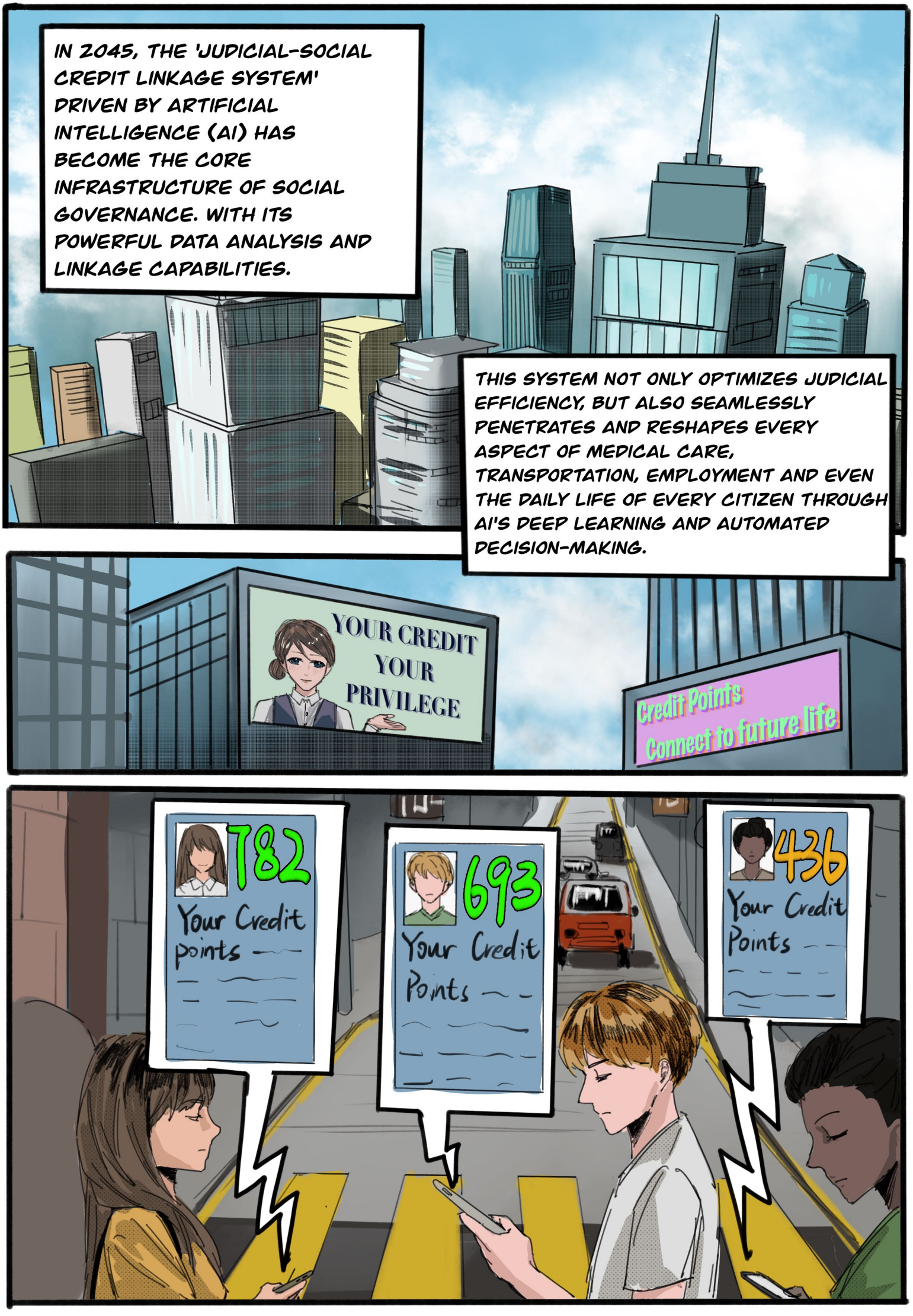
Author: Xiaolu Yi

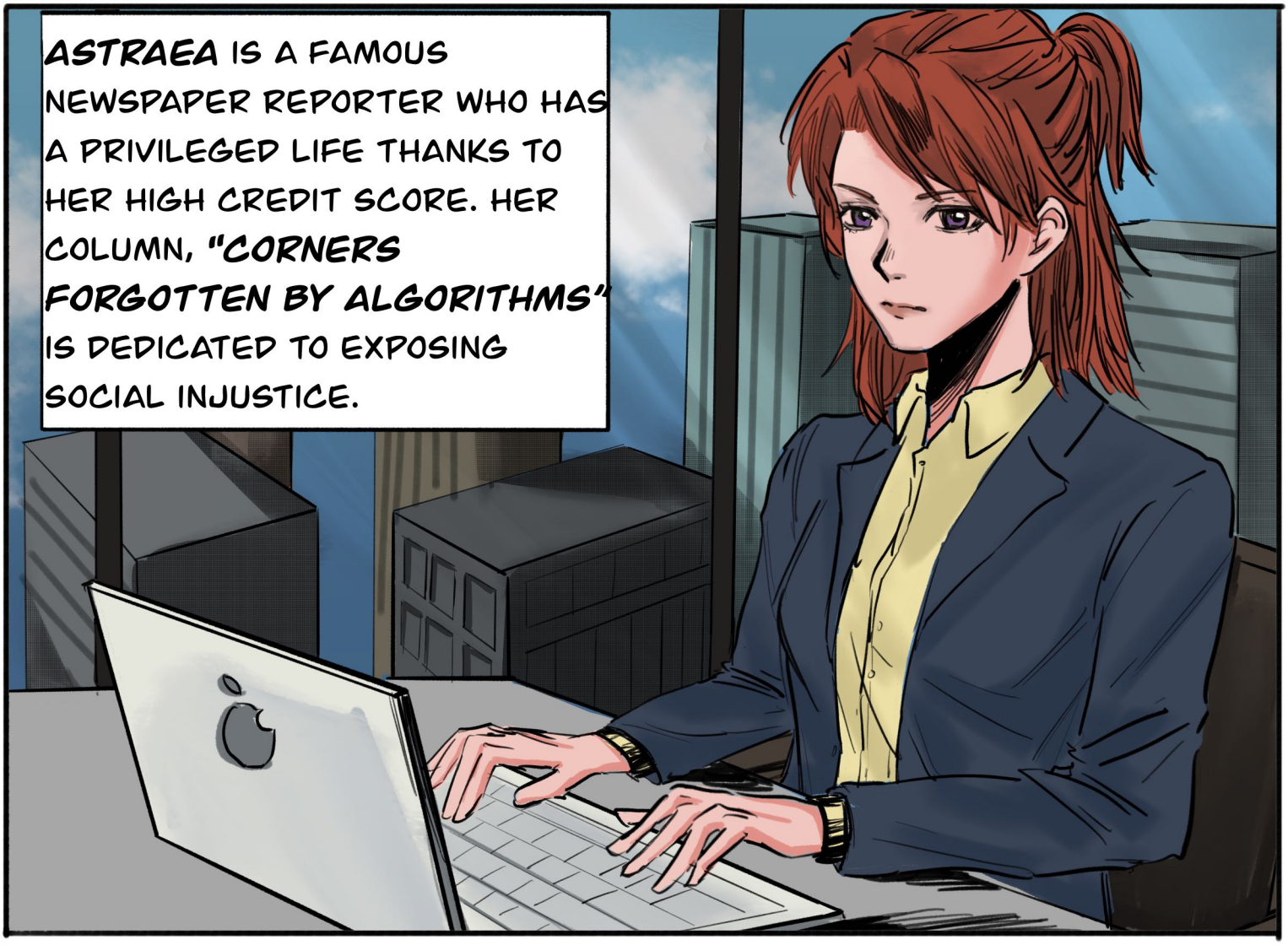
Supervisors: Jan N. van Rijn, Przemyslaw Biecek, Unggul Karami, Francien Dechesne

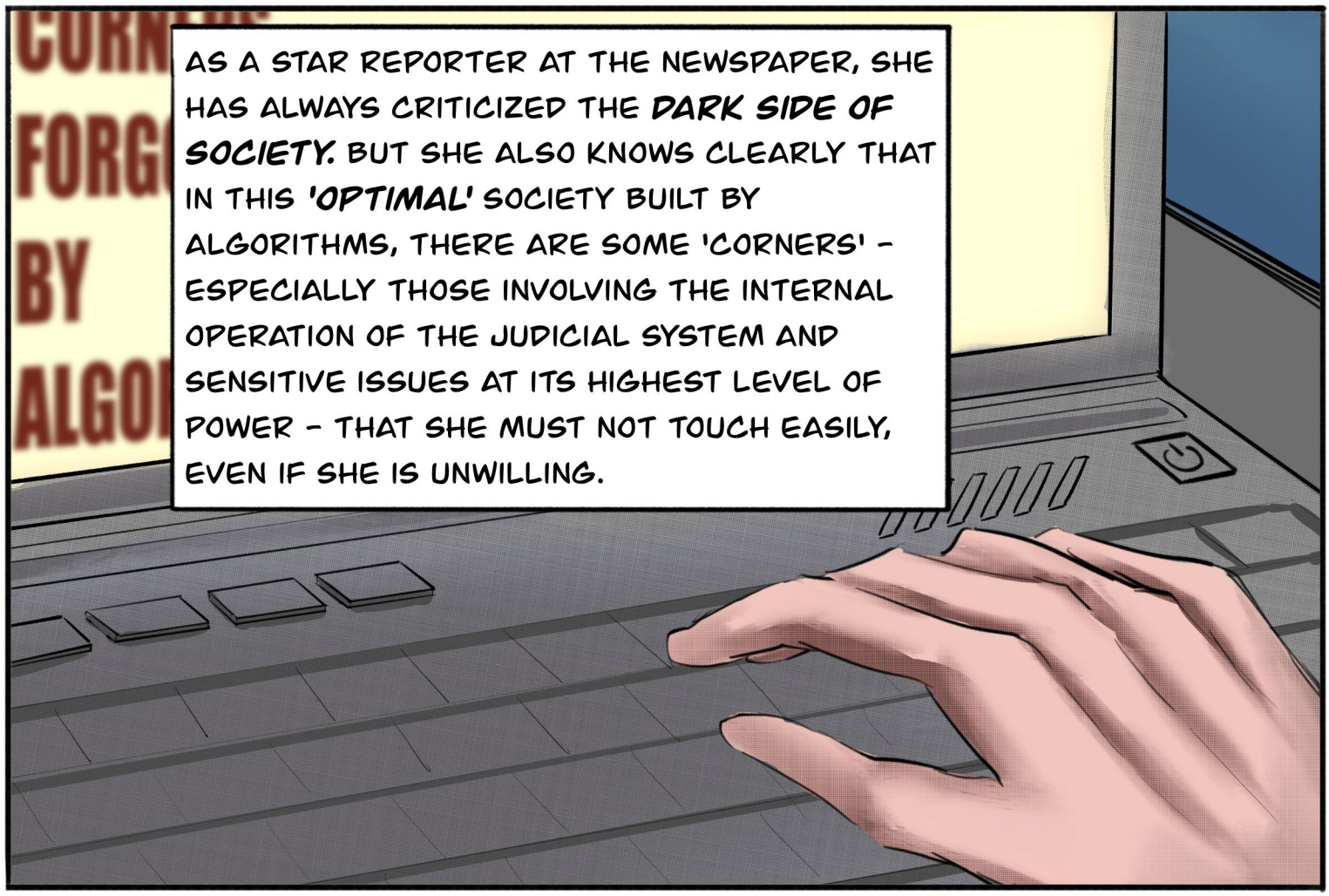
Date: 17th August, 2025

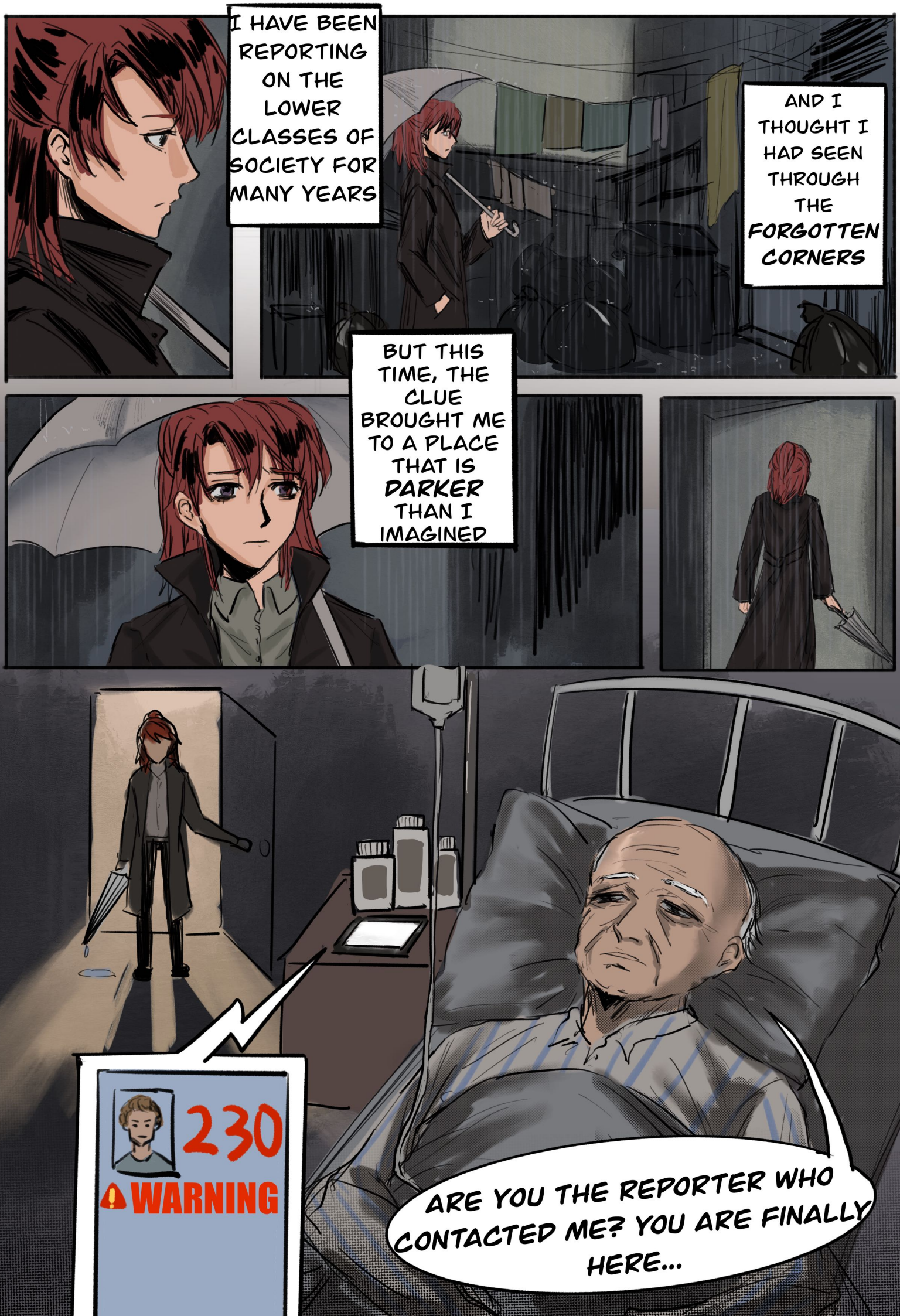
Disclaimer

This is a work of fiction. Names, characters, places, and incidents are either the product of the author's imagination or used fictitiously. Any resemblance to actual persons, living or dead, or actual events is purely coincidental.


IN 2045, THE 'JUDICIAL-SOCIAL CREDIT LINKAGE SYSTEM' DRIVEN BY ARTIFICIAL INTELLIGENCE (AI) HAS BECOME THE CORE INFRASTRUCTURE OF SOCIAL GOVERNANCE. WITH ITS POWERFUL DATA ANALYSIS AND LINKAGE CAPABILITIES.
THIS SYSTEM NOT ONLY OPTIMIZES JUDICIAL EFFICIENCY, BUT ALSO SEAMLESSLY PENETRATES AND RESHAPES EVERY ASPECT OF MEDICAL CARE, TRANSPORTATION, EMPLOYMENT AND EVEN THE DAILY LIFE OF EVERY CITIZEN THROUGH AI'S DEEP LEARNING AND AUTOMATED DECISION-MAKING.
YOUR CREDIT YOUR PRIVILEGE
Credit Points Connect to future life
782
Your Credit points
693
Your Credit Points
436
Your Credit Points


ASTRAEA IS A FAMOUS NEWSPAPER REPORTER WHO HAS A PRIVILEGED LIFE THANKS TO HER HIGH CREDIT SCORE. HER COLUMN, **"CORNERS FORGOTTEN BY ALGORITHMS"** IS DEDICATED TO EXPOSING SOCIAL INJUSTICE.


AS A STAR REPORTER AT THE NEWSPAPER, SHE HAS ALWAYS CRITICIZED THE **DARK SIDE OF SOCIETY.** BUT SHE ALSO KNOWS CLEARLY THAT IN THIS **'OPTIMAL'** SOCIETY BUILT BY ALGORITHMS, THERE ARE SOME 'CORNERS' - ESPECIALLY THOSE INVOLVING THE INTERNAL OPERATION OF THE JUDICIAL SYSTEM AND SENSITIVE ISSUES AT ITS HIGHEST LEVEL OF POWER - THAT SHE MUST NOT TOUCH EASILY, EVEN IF SHE IS UNWILLING.


I HAVE BEEN REPORTING ON THE LOWER CLASSES OF SOCIETY FOR MANY YEARS
AND I THOUGHT I HAD SEEN THROUGH THE FORGOTTEN CORNERS
BUT THIS TIME, THE CLUE BROUGHT ME TO A PLACE THAT IS DARKER THAN I IMAGINED
230
WARNING
ARE YOU THE REPORTER WHO CONTACTED ME? YOU ARE FINALLY HERE...


Warning: Abuse of medical resources!"
LIFE-SAVING MEDICAL TREATMENT WAS DEFINED AS "ABUSE" UNDER THE COLD LOGIC OF THE ALGORITHM. THIS DIRECTLY DESTROYED EVERYTHING THAT SUSTAINED HIS LIVELIHOOD AND DIGNITY.
FOUR CHEMOTHERAPY SESSIONS A MONTH... THIS IS CLEARLY A LIFE-SAVING TREATMENT! HOW COULD IT BECOME... A MEDICAL RESOURCE ABUSER?!
SO... ALL YOUR MEDICAL COVERAGE HAS BEEN CANCELED?
THE SYSTEM SAID THAT I... I TOOK UP TOO MANY PUBLIC RESOURCES... MY CREDIT POINTS... IS GONE...
.....YES.
Pain should not be priced by algorithms - an old man's "death of credit"
I WON'T AVOID IT ANYMORE. THIS TIME, I WILL EXPOSE THE TRUE FACE OF THIS COLD ALGORITHM!
PUBLISH
CLICK!!

Legal warning letter: Infringement notice for false reports

Your Social Credit Score dropped to 635.

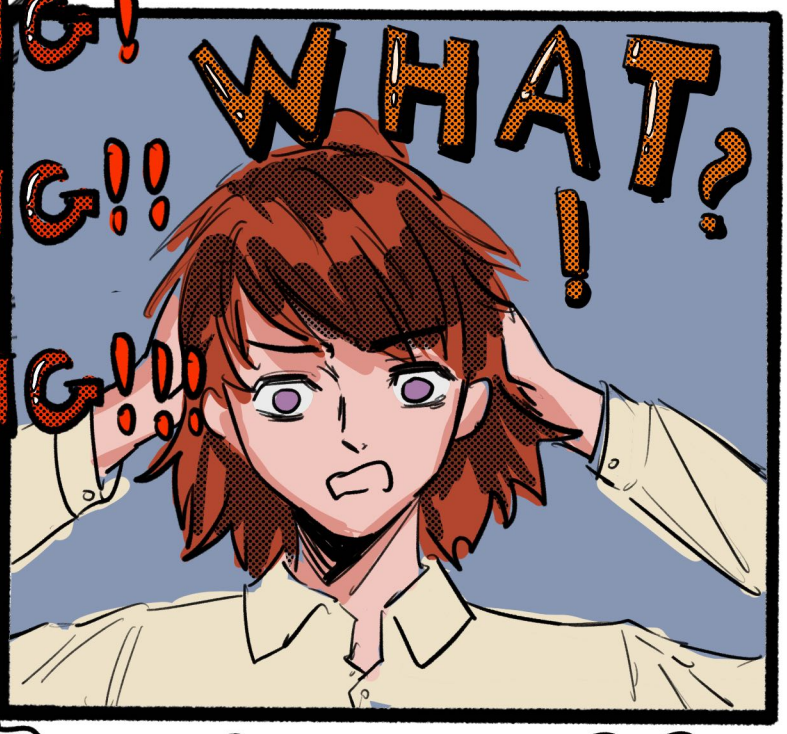
Important Notice: Adjustment of your advanced media permissions

Based on your latest social credit assessment, your advanced news interview permissions and exclusive publishing channels will be temporarily frozen.

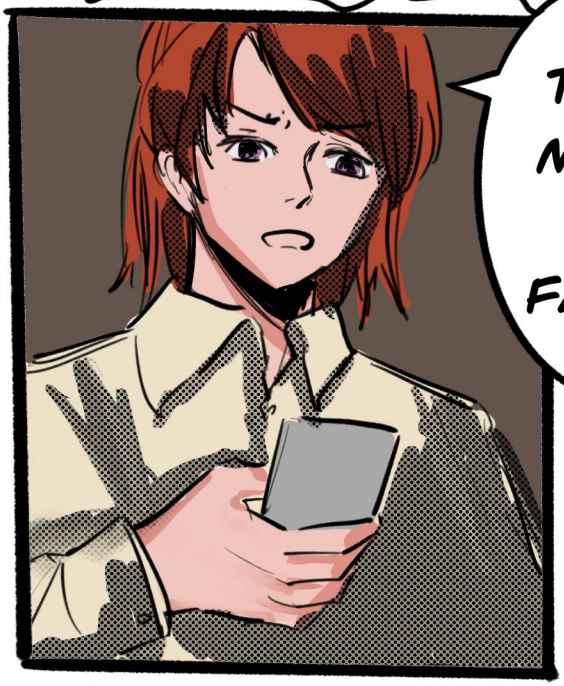
DING!

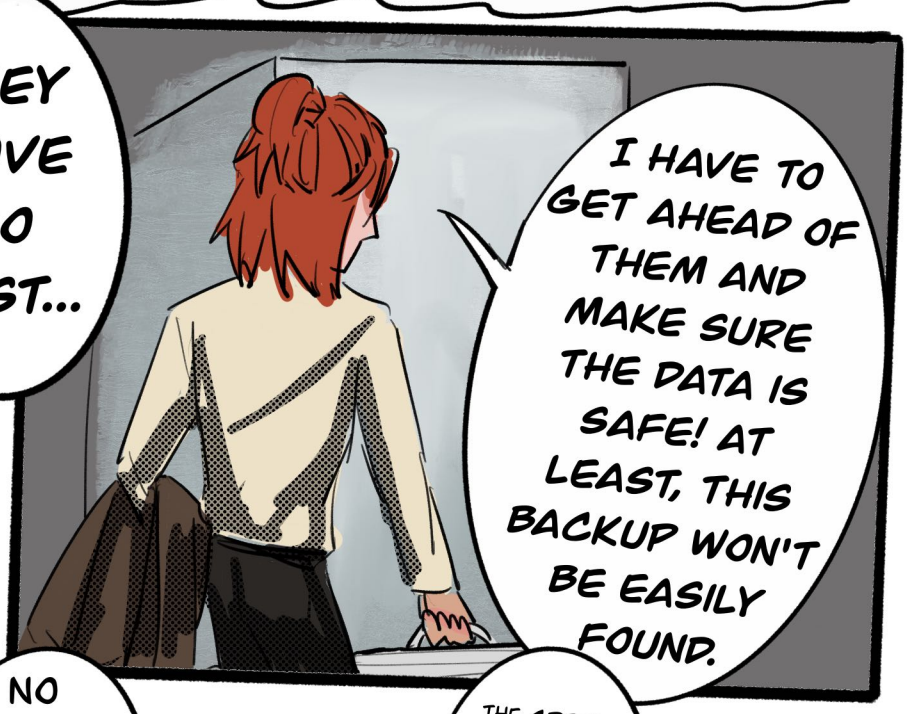
DING!!

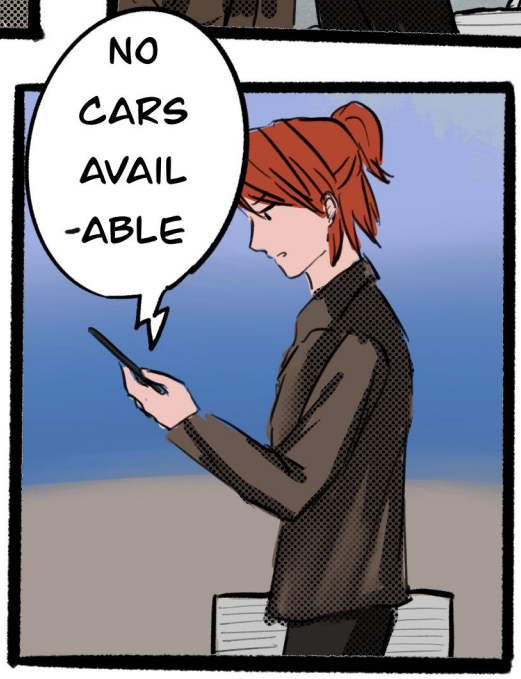
DING!!!

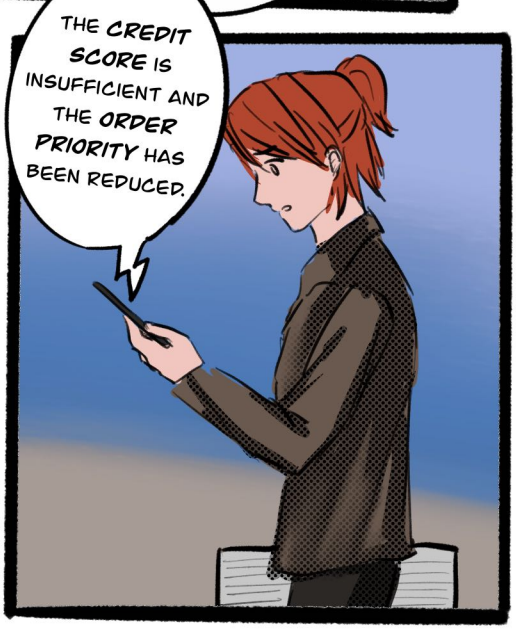


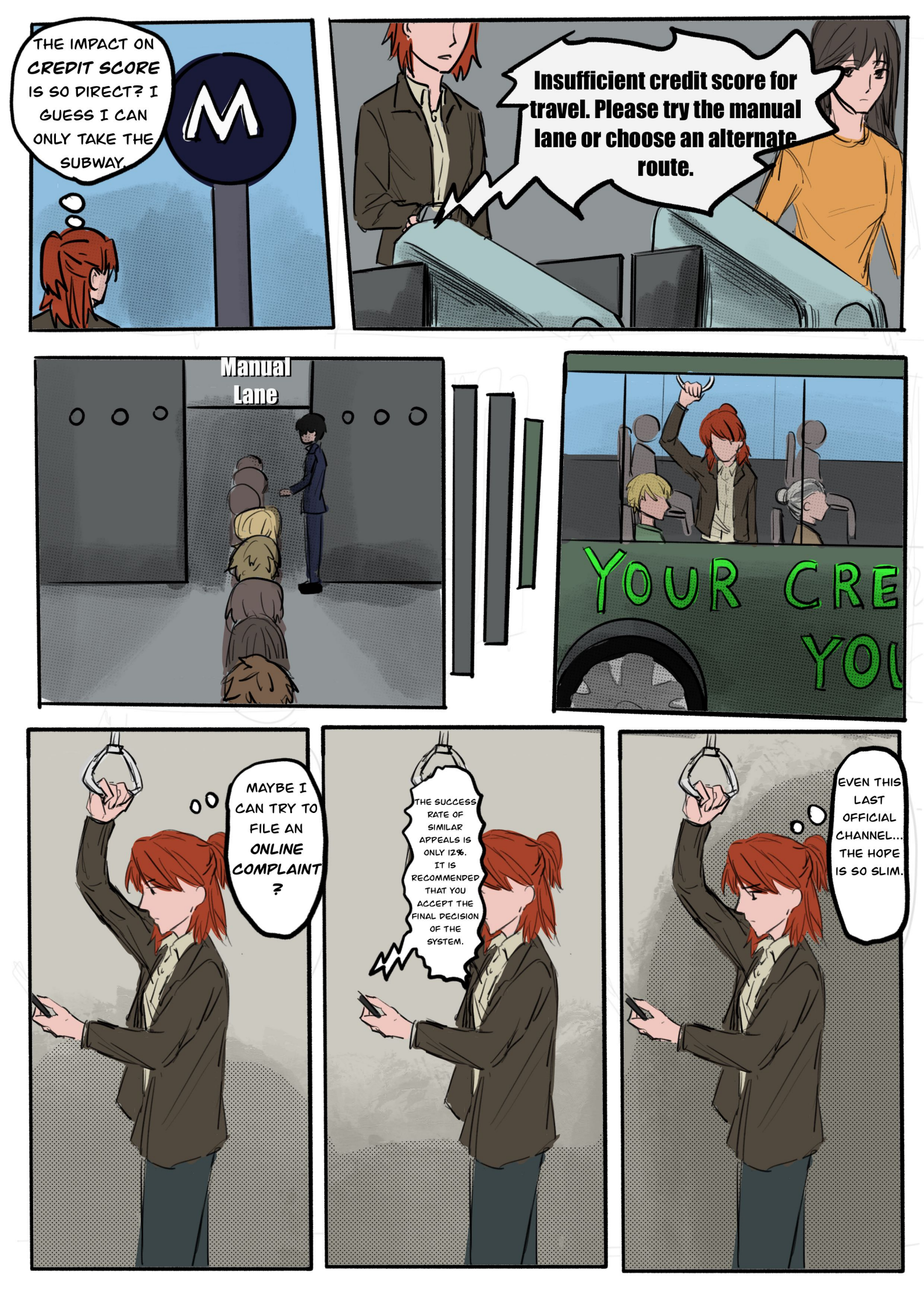
Alert: Credit score has dropped abnormally! Please pay attention to the impact of your personal behavior on your social credit rating!

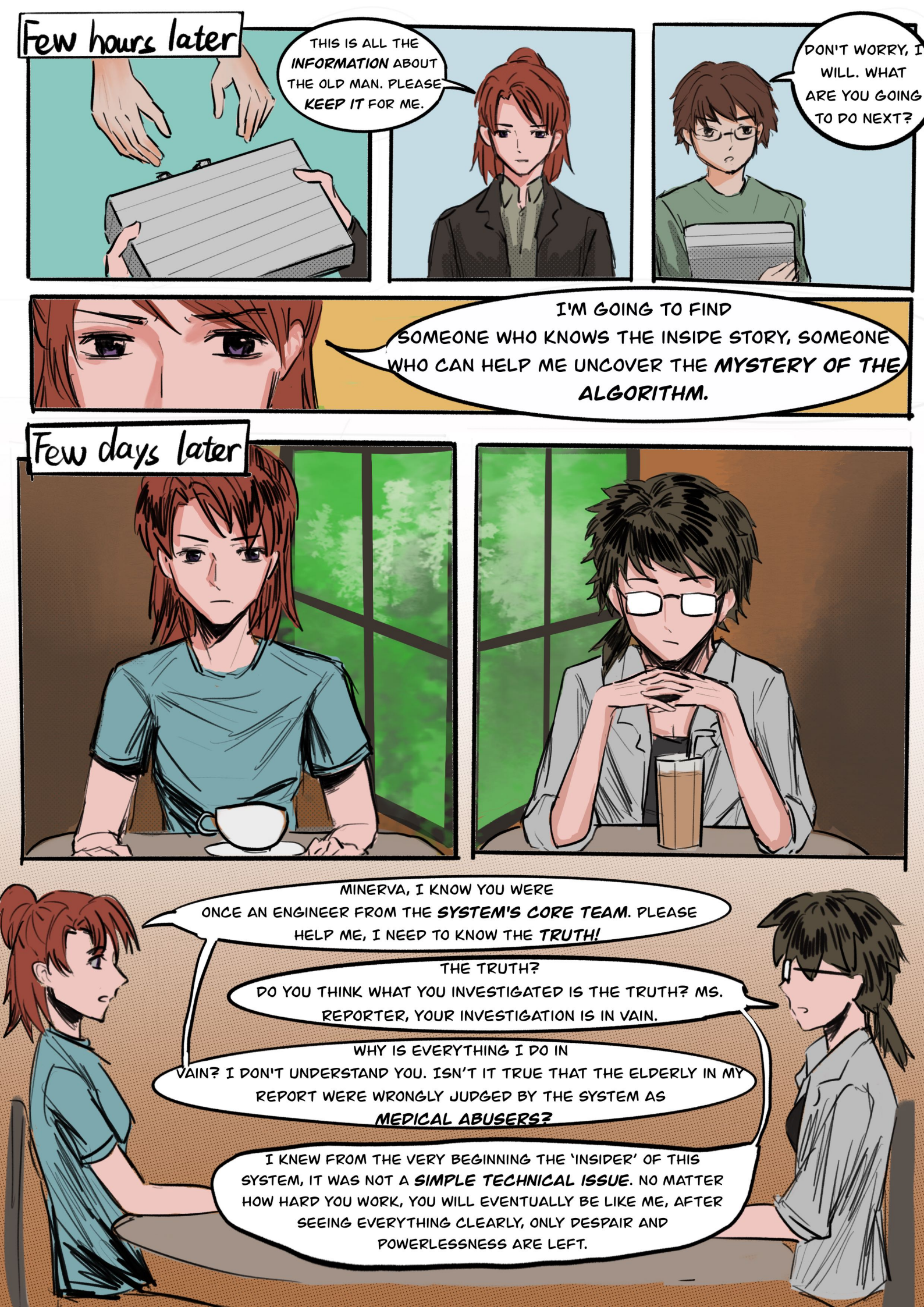


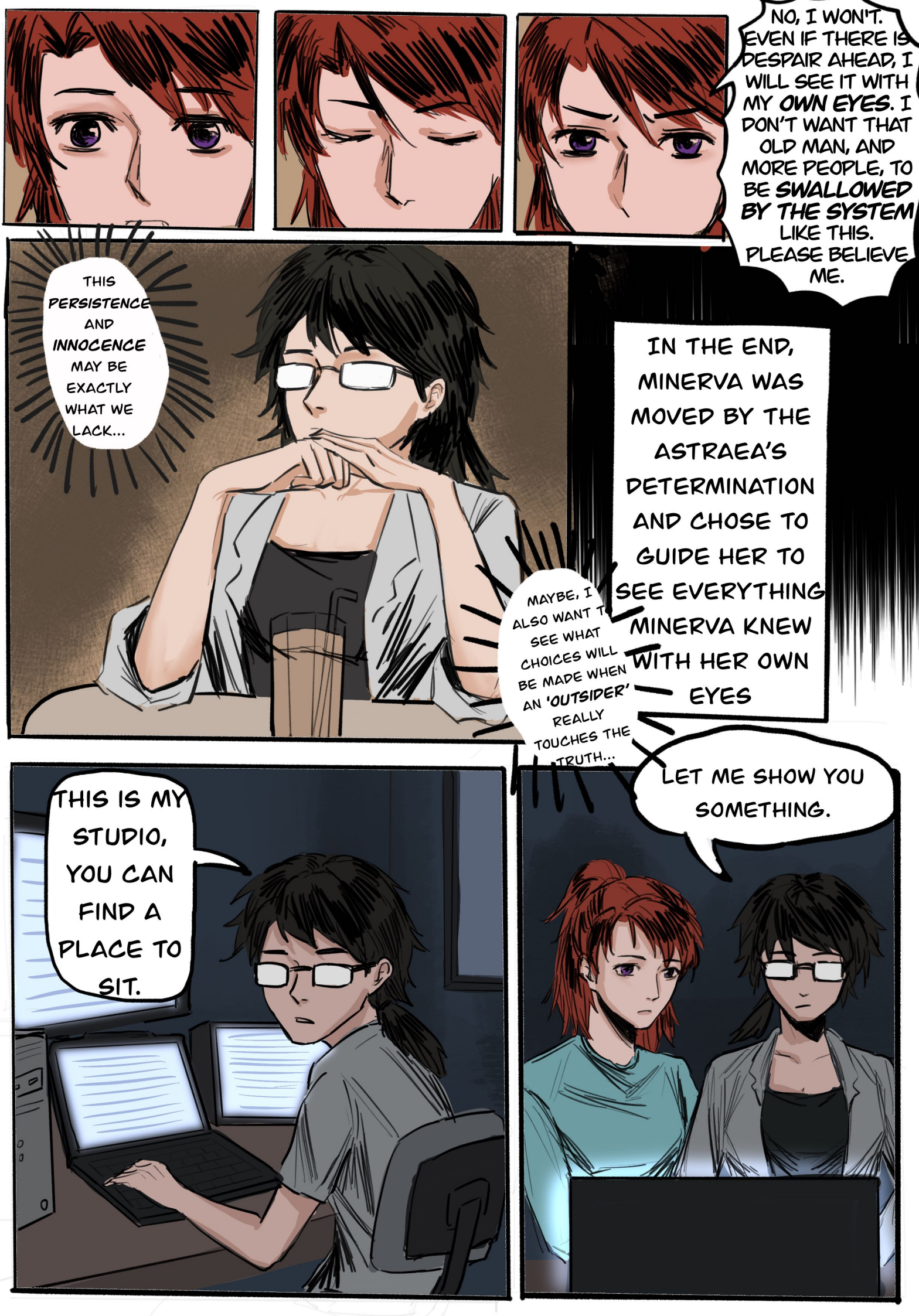


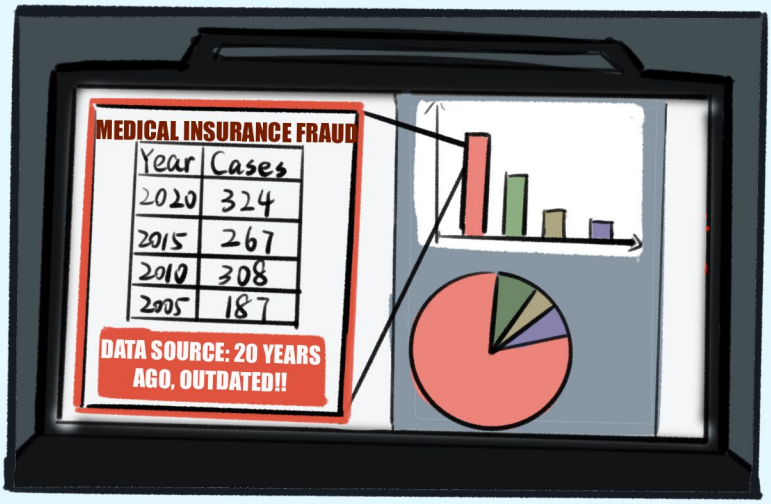


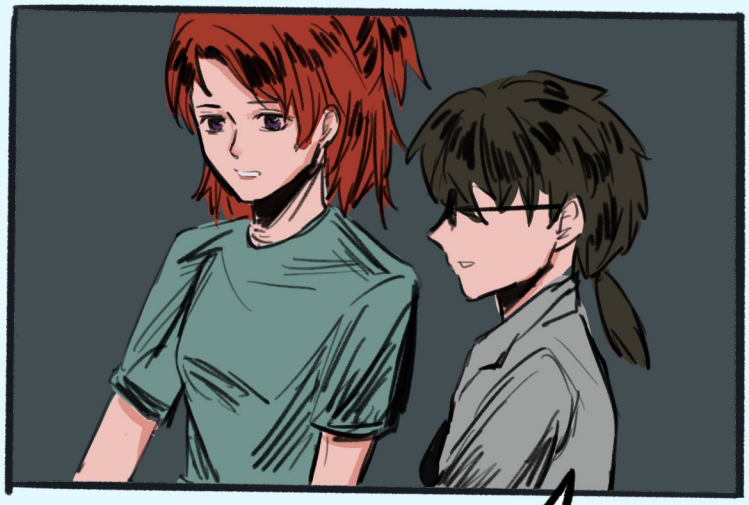



THE IMPACT ON CREDIT SCORE IS SO DIRECT? I GUESS I CAN ONLY TAKE THE SUBWAY.
M
Insufficient credit score for travel. Please try the manual lane or choose an alternate route.
Manual Lane
YOUR CRE
YOU
MAYBE I CAN TRY TO FILE AN ONLINE COMPLAINT ?
THE SUCCESS RATE OF SIMILAR APPEALS IS ONLY 12%. IT IS RECOMMENDED THAT YOU ACCEPT THE FINAL DECISION OF THE SYSTEM.
EVEN THIS LAST OFFICIAL CHANNEL... THE HOPE IS SO SLIM.


Few hours later
THIS IS ALL THE INFORMATION ABOUT THE OLD MAN. PLEASE KEEP IT FOR ME.
DON'T WORRY, I WILL. WHAT ARE YOU GOING TO DO NEXT?
I'M GOING TO FIND SOMEONE WHO KNOWS THE INSIDE STORY, SOMEONE WHO CAN HELP ME UNCOVER THE MYSTERY OF THE ALGORITHM.
Few days later
MINERVA, I KNOW YOU WERE ONCE AN ENGINEER FROM THE SYSTEM'S CORE TEAM. PLEASE HELP ME, I NEED TO KNOW THE TRUTH!
THE TRUTH?
DO YOU THINK WHAT YOU INVESTIGATED IS THE TRUTH? MS. REPORTER, YOUR INVESTIGATION IS IN VAIN.
WHY IS EVERYTHING I DO IN VAIN? I DON'T UNDERSTAND YOU. ISN'T IT TRUE THAT THE ELDERLY IN MY REPORT WERE WRONGLY JUDGED BY THE SYSTEM AS MEDICAL ABUSERS?
I KNEW FROM THE VERY BEGINNING THE 'INSIDER' OF THIS SYSTEM, IT WAS NOT A SIMPLE TECHNICAL ISSUE. NO MATTER HOW HARD YOU WORK, YOU WILL EVENTUALLY BE LIKE ME, AFTER SEEING EVERYTHING CLEARLY, ONLY DESPAIR AND POWERLESSNESS ARE LEFT.


NO, I WON'T. EVEN IF THERE IS DESPAIR AHEAD, I WILL SEE IT WITH MY OWN EYES. I DON'T WANT THAT OLD MAN, AND MORE PEOPLE, TO BE SWALLOWED BY THE SYSTEM LIKE THIS. PLEASE BELIEVE ME.
THIS PERSISTENCE AND INNOCENCE MAY BE EXACTLY WHAT WE LACK...
IN THE END, MINERVA WAS MOVED BY THE ASTRAEA'S DETERMINATION AND CHOSE TO GUIDE HER TO SEE EVERYTHING MINERVA KNEW WITH HER OWN EYES
MAYBE, I ALSO WANT TO SEE WHAT CHOICES WILL BE MADE WHEN AN 'OUTSIDER' REALLY TOUCHES THE TRUTH...
LET ME SHOW YOU SOMETHING.
THIS IS MY STUDIO, YOU CAN FIND A PLACE TO SIT.





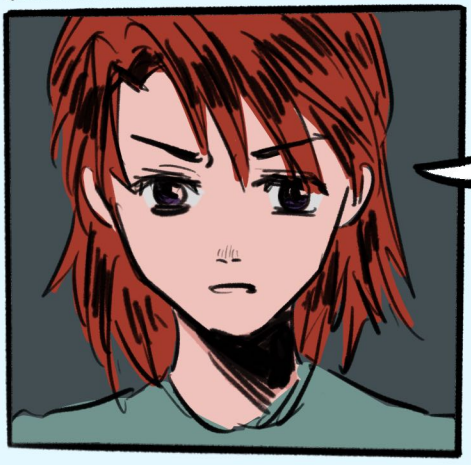
LOOK HERE... THE ALGORITHMS IN THE MEDICAL PART ARE STILL OVERLY RELYING ON ***MEDICAL INSURANCE FRAUD DATA 20 YEARS AGO.*** THE MEDICAL MODEL AT THAT TIME WAS ***COMPLETELY DIFFERENT*** FROM NOW, AND THIS DATA SHOULD HAVE BEEN ***ELIMINATED*** LONG AGO.

WAIT... I STILL DON'T UNDERSTAND. EVEN IF THE DATA IS OUTDATED, HOW CAN THE SYSTEM ***DIRECTLY*** COME TO THE CONCLUSION OF "ABUSE OF MEDICAL RESOURCES"? HOW DOES THE ***PROCESS WORK IN BETWEEN?***

YOU ASKED A GOOD QUESTION, WHICH IS WHAT WE OFTEN CALL THE AI '***BLACK BOX***' PROBLEM. SIMPLY PUT, IT REFERS TO THE FACT THAT THE ***INTERNAL DECISION-MAKING PROCESS*** OF SOME HIGHLY COMPLEX AI SYSTEMS, ESPECIALLY DEEP LEARNING MODELS, IS COMPLETELY ***OPAQUE AND UNEXPLAINABLE*** TO HUMANS.
WE INPUT DATA INTO IT AND IT WILL GIVE US THE RESULT, BUT WE CANNOT FOLLOW AND UNDERSTAND HOW IT ***REASONED, WEIGHED AND FINALLY REACHED THE RESULT*** STEP BY STEP LIKE TRADITIONAL PROGRAMS. TO USE A COMMON METAPHOR, IT IS LIKE AN ***"OPAQUE BOX".*** WE ONLY KNOW WHAT WE THROW IN AND WHAT WE GET, BUT WE HAVE NO IDEA WHAT HAPPENS ***INSIDE THE BOX.***

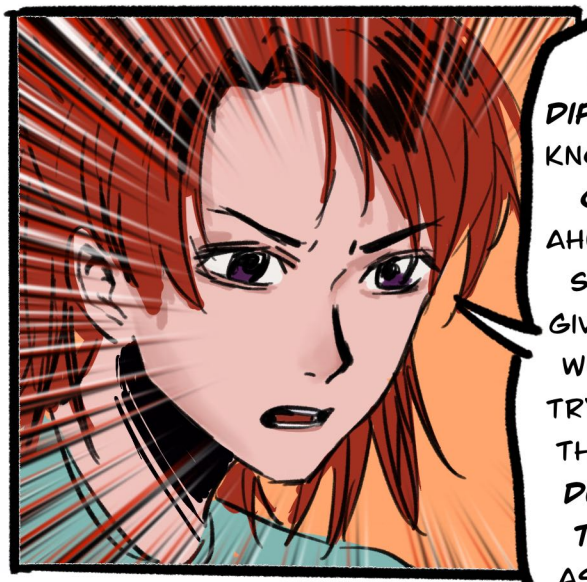
FOR THE PROBLEMS WE ENCOUNTER, SUCH AS ***OUTDATED MEDICAL INSURANCE DATA,*** AFTER IT IS ABSORBED BY AI, THE SYSTEM WILL USE IT TO TRAIN THE MODEL, AND THEN COMBINE IT WITH MILLIONS OF OTHER INVISIBLE ***VARIABLES AND WEIGHTS,*** AND FINALLY FORM A JUDGMENT IN THE 'BLACK BOX' THAT WE CANNOT UNDERSTAND. WE CAN ONLY SEE THE DATA ***INPUT*** (OLD DATA) AND THE RESULT ***OUTPUT*** (MISJUDGMENT), BUT WE CANNOT KNOW ***HOW*** THIS 'BLACK BOX' CONNECTS THE TWO.

I SEE... SO, IT IS NOT ONLY A TECHNICAL PROBLEM, BUT ALSO A PIECE OF CLOTH TO HIDE THE BEHIND-THE-SCENES OPERATIONS, MAKING THOSE ***INJUSTICES IMPECCABLE.***



SINCE YOU CAN DIG UP THESE ***EVIDENCES,*** THERE MUST BE OTHERS WHO CAN. WHY HAS NO ONE ***EXPOSED*** THESE THINGS BEFORE?

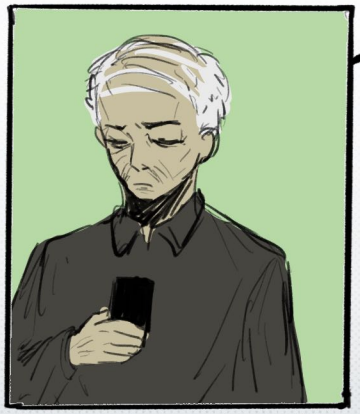
EXPOSE?
MANY PEOPLE HAVE TRIED, BUT IT'S TOO DIFFICULT. DO YOU THINK THOSE WHO KNOW THE TRUTH WILL ***ALLOW THE EVIDENCE TO BE EXPOSED?*** SOME FORCES ARE FAR MORE ***POWERFUL*** THAN YOU THINK...



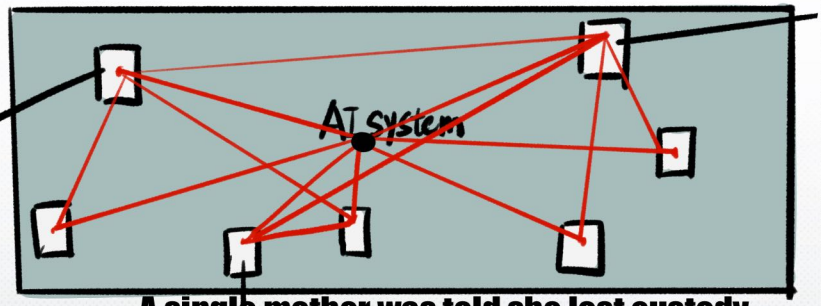
I KNOW IT'S ***DIFFICULT,*** AND I KNOW THERE ARE ***OBSTACLES*** AHEAD. BUT EVEN SO, I WANT TO GIVE IT A TRY. IF WE DON'T EVEN TRY, WHO SHOULD THOSE WHO ARE ***DEVOURED BY THE SYSTEM*** ASK FOR HELP?

YOU ARE SUCH A ***NAIVE AND STUBBORN*** GUY. BUT PERHAPS, THIS IS EXACTLY WHAT THIS QUAGMIRE NEEDS. OK, ASTRAEA, I'LL BE WITH YOU. HOWEVER, ONCE IT STARTS, THERE IS ***NO TURNING BACK.***

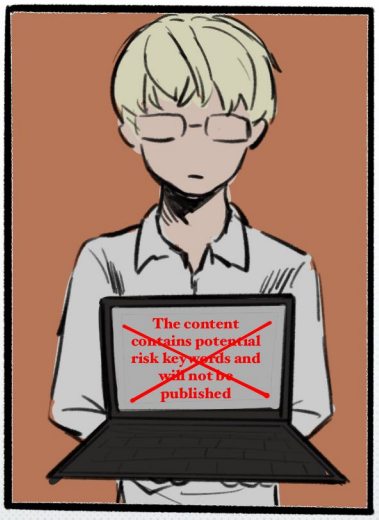
Few Weeks later



A veteran was banned by a takeaway platform because of his comrades' criminal record



A single mother was told she lost custody of her children due to her ex-husband's debt problems

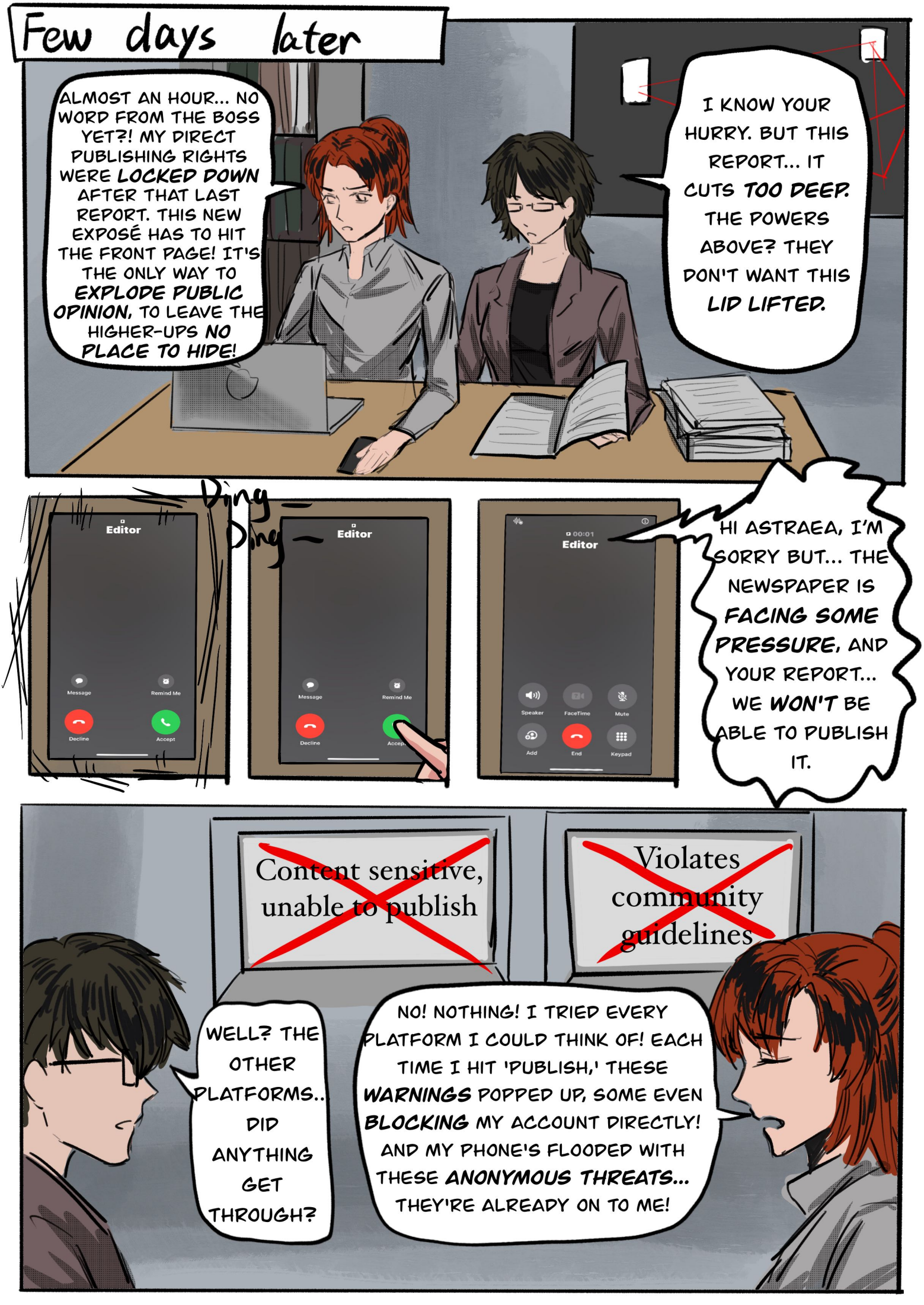


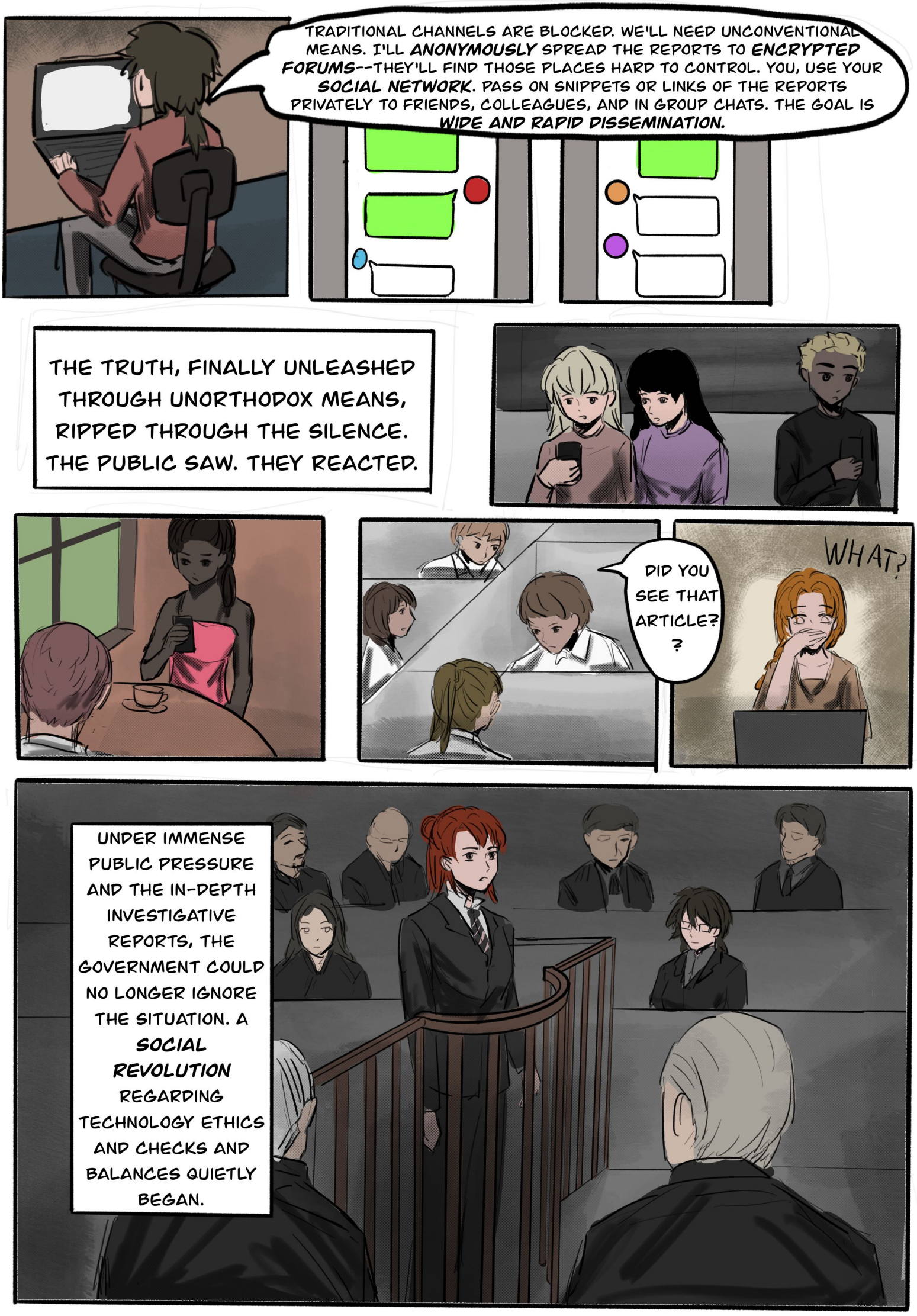
A university student's academic paper was judged by "academic AI" to contain "dangerous ideas" and cannot be published

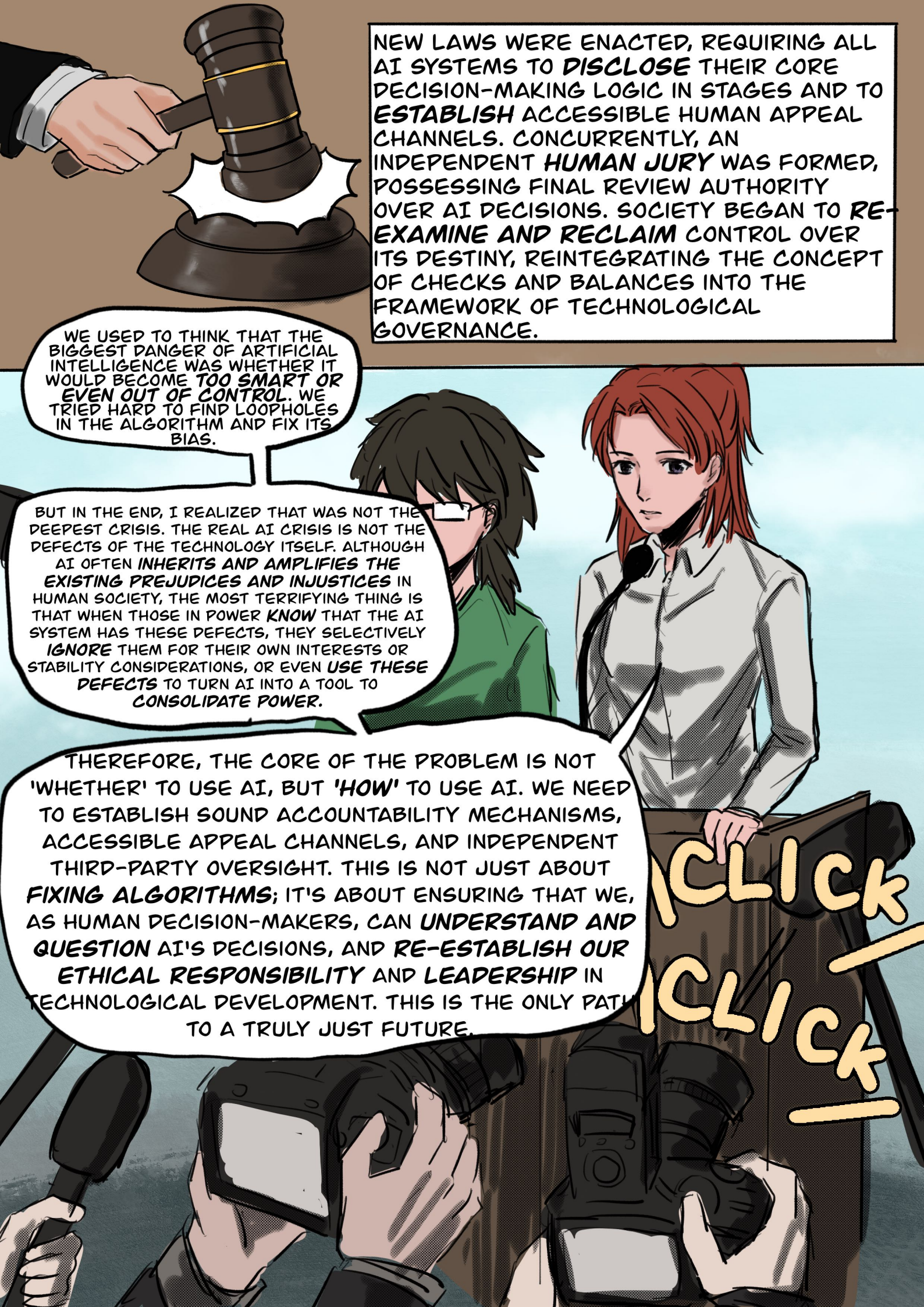
SO, THESE PEOPLE THEMSELVES ARE VICTIMS OF ***EXISTING SOCIETAL BIAS,*** AND THE AI JUST ***LEARNED AND AMPLIFIED*** THESE BIASES.....

THAT'S RIGHT. WE HAVE FOUND SO MUCH EVIDENCE THAT THESE FLAWS ARE ***NOT ACCIDENTAL,*** BUT RATHER ***INTENTIONAL*** IN SYSTEM DESIGN AND DECISION-MAKING.

THESE CASES ARE THE REAL CONSEQUENCES OF ALGORITHMIC BIAS. THESE CORNERS THAT WERE ***'FORGOTTEN'*** BY THE SYSTEM TELL A STORY DESTROYED BY ALGORITHMIC BIASES. THESE TRAGIC STORIES ARE THE MOST POWERFUL ***SOCIAL EVIDENCES*** TO REVEAL THE TRUTH.


Few days later
ALMOST AN HOUR... NO WORD FROM THE BOSS YET?! MY DIRECT PUBLISHING RIGHTS WERE LOCKED DOWN AFTER THAT LAST REPORT. THIS NEW EXPOSÉ HAS TO HIT THE FRONT PAGE! IT'S THE ONLY WAY TO EXPLODE PUBLIC OPINION, TO LEAVE THE HIGHER-UPS NO PLACE TO HIDE!
I KNOW YOUR HURRY. BUT THIS REPORT... IT CUTS TOO DEEP. THE POWERS ABOVE? THEY DON'T WANT THIS LID LIFTED.
Ding—
Ding—
Editor
Message
Remind Me
Decline
Accept
00:01
Speaker
FaceTime
Mute
Add
End
Keypad
HI ASTRAEA, I'M SORRY BUT... THE NEWSPAPER IS FACING SOME PRESSURE, AND YOUR REPORT... WE WON'T BE ABLE TO PUBLISH IT.
Content sensitive, unable to publish
Violates community guidelines
WELL? THE OTHER PLATFORMS.. DID ANYTHING GET THROUGH?
NO! NOTHING! I TRIED EVERY PLATFORM I COULD THINK OF! EACH TIME I HIT 'PUBLISH,' THESE WARNINGS POPPED UP, SOME EVEN BLOCKING MY ACCOUNT DIRECTLY! AND MY PHONE'S FLOODED WITH THESE ANONYMOUS THREATS... THEY'RE ALREADY ON TO ME!


TRADITIONAL CHANNELS ARE BLOCKED. WE'LL NEED UNCONVENTIONAL MEANS. I'LL ANONYMOUSLY SPREAD THE REPORTS TO ENCRYPTED FORUMS--THEY'LL FIND THOSE PLACES HARD TO CONTROL. YOU, USE YOUR SOCIAL NETWORK. PASS ON SNIPPETS OR LINKS OF THE REPORTS PRIVATELY TO FRIENDS, COLLEAGUES, AND IN GROUP CHATS. THE GOAL IS WIDE AND RAPID DISSEMINATION.
THE TRUTH, FINALLY UNLEASHED THROUGH UNORTHODOX MEANS, RIPPED THROUGH THE SILENCE. THE PUBLIC SAW. THEY REACTED.
DID YOU SEE THAT ARTICLE??
WHAT?
UNDER IMMENSE PUBLIC PRESSURE AND THE IN-DEPTH INVESTIGATIVE REPORTS, THE GOVERNMENT COULD NO LONGER IGNORE THE SITUATION. A SOCIAL REVOLUTION REGARDING TECHNOLOGY ETHICS AND CHECKS AND BALANCES QUIETLY BEGAN.


NEW LAWS WERE ENACTED, REQUIRING ALL AI SYSTEMS TO DISCLOSE THEIR CORE DECISION-MAKING LOGIC IN STAGES AND TO ESTABLISH ACCESSIBLE HUMAN APPEAL CHANNELS. CONCURRENTLY, AN INDEPENDENT HUMAN JURY WAS FORMED, POSSESSING FINAL REVIEW AUTHORITY OVER AI DECISIONS. SOCIETY BEGAN TO RE-EXAMINE AND RECLAIM CONTROL OVER ITS DESTINY, REINTEGRATING THE CONCEPT OF CHECKS AND BALANCES INTO THE FRAMEWORK OF TECHNOLOGICAL GOVERNANCE.
WE USED TO THINK THAT THE BIGGEST DANGER OF ARTIFICIAL INTELLIGENCE WAS WHETHER IT WOULD BECOME TOO SMART OR EVEN OUT OF CONTROL. WE TRIED HARD TO FIND LOOPHOLES IN THE ALGORITHM AND FIX ITS BIAS.
BUT IN THE END, I REALIZED THAT WAS NOT THE DEEPEST CRISIS. THE REAL AI CRISIS IS NOT THE DEFECTS OF THE TECHNOLOGY ITSELF. ALTHOUGH AI OFTEN INHERITS AND AMPLIFIES THE EXISTING PREJUDICES AND INJUSTICES IN HUMAN SOCIETY, THE MOST TERRIFYING THING IS THAT WHEN THOSE IN POWER KNOW THAT THE AI SYSTEM HAS THESE DEFECTS, THEY SELECTIVELY IGNORE THEM FOR THEIR OWN INTERESTS OR STABILITY CONSIDERATIONS, OR EVEN USE THESE DEFECTS TO TURN AI INTO A TOOL TO CONSOLIDATE POWER.
THEREFORE, THE CORE OF THE PROBLEM IS NOT 'WHETHER' TO USE AI, BUT 'HOW' TO USE AI. WE NEED TO ESTABLISH SOUND ACCOUNTABILITY MECHANISMS, ACCESSIBLE APPEAL CHANNELS, AND INDEPENDENT THIRD-PARTY OVERSIGHT. THIS IS NOT JUST ABOUT FIXING ALGORITHMS; IT'S ABOUT ENSURING THAT WE, AS HUMAN DECISION-MAKERS, CAN UNDERSTAND AND QUESTION AI'S DECISIONS, AND RE-ESTABLISH OUR ETHICAL RESPONSIBILITY AND LEADERSHIP IN TECHNOLOGICAL DEVELOPMENT. THIS IS THE ONLY PATH TO A TRULY JUST FUTURE.
CLICK
CLICK

Reference List:

Hoffmann-Riem, W. (2020). Artificial Intelligence as a Challenge for Law and Regulation. In: Wischmeyer, T., Rademacher, T. (eds) Regulating Artificial Intelligence. Springer, Cham. https://doi.org/10.1007/978-3-030-32361-5_1

S. Raaijmakers, "Artificial Intelligence for Law Enforcement: Challenges and Opportunities," in IEEE Security & Privacy, vol. 17, no. 5, pp. 74-77, Sept.-Oct. 2019, doi: 10.1109/MSEC.2019.2925649.

Hildebrandt, Mireille, Law As Computation in the Era of Artificial Legal Intelligence. Speaking Law to the Power of Statistics (June 7, 2017). Available at SSRN: https://ssrn.com/abstract=2983045 or http://dx.doi.org/10.2139/ssrn.2983045

Petit, Nicolas, Law and Regulation of Artificial Intelligence and Robots - Conceptual Framework and Normative Implications (March 9, 2017). Available at SSRN: https://ssrn.com/abstract=2931339 or http://dx.doi.org/10.2139/ssrn.2931339

W. Bradley Wendel, The Promise and Limitations of Artificial Intelligence in the Practice of Law, 72 Okla. L. Rev. 21 (2019).

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. Telecommunications Policy, 44(6).

Greenstein, S. Preserving the rule of law in the era of artificial intelligence (AI). Artif Intell Law 30, 291–323 (2022). https://doi.org/10.1007/s10506-021-09294-4

Alarie, B., Niblett, A., & Yoon, A. H. (2018). How artificial intelligence will affect the practice of law. University of Toronto Law Journal, 68(supplement 1), 106-124.

Richmond, K.M., Muddamsetty, S.M., Gammeltoft-Hansen, T. et al. Explainable AI and Law: An Evidential Survey. DISO 3, 1 (2024). https://doi.org/10.1007/s44206-023-00081-z