

# Accountable Algorithms

Edward W. Felten

Joshua Kroll

Center for Information Technology Policy  
Dept. of Computer Science  
Princeton University

Fairness, Accountability and Transparency in ML Workshop, NIPS 2014  
December 12, 2014

Accountability != Transparency

# **Example: Tax Audit**

# Example: Tax Audit



subject

# Example: Tax Audit



subject



agency

## **Example: Tax Audit**



## subject

**1040** **U.S. Individual Income Tax Return 2013**

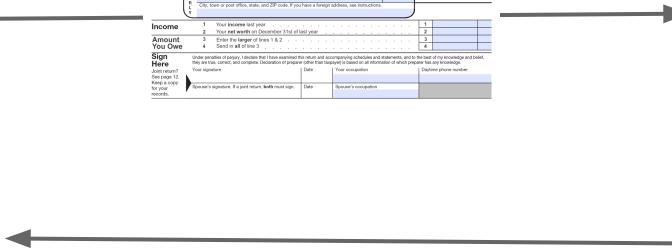
(See instructions on reverse side for filing by mail or online.)

Name, Address, and SSN  See separate page.	Phone number  Last name  First name  Middle initial  Address number and street or box, if you have a P.O. box, see instructions.  Apartment, suite, or room number  City, town, or post office, state and zip code. If you have a P.O. box, see instructions.	Social Security number  Your social security number  Sponsor's social security number
Income  Amount You Give	1 Your Known and prior 2 Your net worth as of December 31st of last year 3 Send it in a separate envelope 4	1 2 3 4
Sign  My signature for joint return or my spouse's signature for separate return  Spouse's signature if a joint return, both must sign  Spouse's signature if a separate return	My declaration I declare under penalty of perjury that I have read the return and accompanying schedules and statements, and to the best of my knowledge and belief, the information contained in this return is true, correct, and complete. I understand that any false statement or omission may subject me to penalties and criminal prosecution.  Signature Date  Signature Date  Signature Date	My declaration I declare under penalty of perjury that I have read the return and accompanying schedules and statements, and to the best of my knowledge and belief, the information contained in this return is true, correct, and complete. I understand that any false statement or omission may subject me to penalties and criminal prosecution.  Signature Date  Signature Date  Signature Date



## agency

# Example: Tax Audit



subject

"We're auditing you."



agency

# **Accountable Tax Audit**



## subject

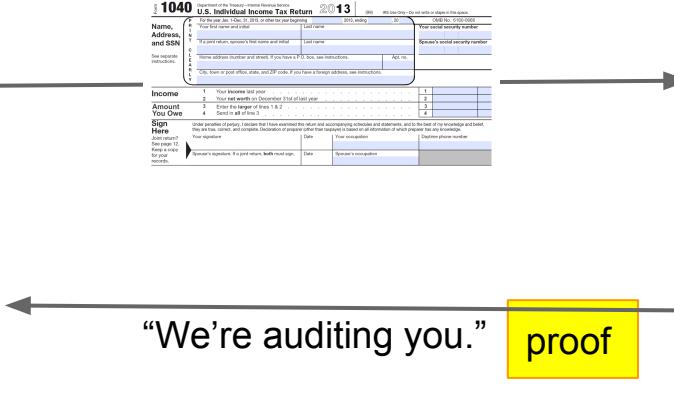
<b>1040</b>		<b>U.S. Individual Income Tax Return</b>	<b>2013</b>	IRS Online - Quick-File or eFile
Name, Address, and SSN  Instructions		OMB No. 1545-0130 Quesada, 2013		
<input type="checkbox"/> I am filing this 1040 as my tax return for the year ending <u>2013</u> . <input type="checkbox"/> I am filing this 1040 as my tax return for the year ending <u>2012</u> .		Your social security number Spouse's social security number		
<input type="checkbox"/> I am a citizen. <input type="checkbox"/> I am a resident alien. <input type="checkbox"/> I am a nonresident alien. <input type="checkbox"/> I am a spouse of a citizen or resident alien. <input type="checkbox"/> I am a dependent of a citizen or resident alien. <input type="checkbox"/> I am a nonresident alien who has filed Form 8815, See Instructions.		Name Last name, first name, middle initial Street address and city, state and ZIP code, see instructions Apt. no.		
<b>Income</b> You have been paid by whom You owe Here See page 12 for our tax return filing date		1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 8010 8011 8012 8013 8014 8015 8016 8017 8018 8019 8020 8021 8022 8023 8024 8025 8026 8027 8028 8029 8030 8031 8032 8033 8034 8035 8036 8037 8038 8039 8040 8041 8042 8043 8044 8045 8046 8047 8048 8049 8050 8051 8052 8053 8054 8055 8056 8057 8058 8059 8060 8061 8062 8063 8064 8065 8066 8067 8068 8069 8070 8071 8072 8073 8074 8075 8076 8077 8078 8079 8080 8081 8082 8083 8084 8085 8086 8087 8088 8089 8090 8091 8092 8093 8094 8095 8096 8097 8098 8099 80100 80101 80102 80103 80104 80105 80106 80107 80108 80109 80110 80111 80112 80113 80114 80115 80116 80117 80118 80119 80120 80121 80122 80123 80124 80125 80126 80127 80128 80129 80130 80131 80132 80133 80134 80135 80136 80137 80138 80139 80140 80141 80142 80143 80144 80145 80146 80147 80148 80149 80150 80151 80152 80153 80154 80155 80156 80157 80158 80159 80160 80161 80162 80163 80164 80165 80166 80167 80168 80169 80170 80171 80172 80173 80174 80175 80176 80177 80178 80179 80180 80181 80182 80183 80184 80185 80186 80187 80188 80189 80190 80191 80192 80193 80194 80195 80196 80197 80198 80199 801000 801001 801002 801003 801004 801005 801006 801007 801008 801009 8010010 8010011 8010012 8010013 8010014 8010015 8010016 8010017 8010018 8010019 80100100 80100101 80100102 80100103 80100104 80100105 80100106 80100107 80100108 80100109 801001000 801001001 801001002 801001003 801001004 801001005 801001006 801001007 801001008 801001009 8010010000 8010010010 8010010020 8010010030 8010010040 8010010050 8010010060 8010010070 8010010080 8010010090 8010010001 8010010011 8010010021 8010010031 8010010041 8010010051 8010010061 8010010071 8010010081 8010010091 8010010002 8010010012 8010010022 8010010032 8010010042 8010010052 8010010062 8010010072 8010010082 8010010092 8010010003 8010010013 8010010023 8010010033 8010010043 8010010053 8010010063 8010010073 8010010083 8010010093 8010010004 8010010014 8010010024 8010010034 8010010044 8010010054 8010010064 8010010074 8010010084 8010010094 8010010005 8010010015 8010010025 8010010035 8010010045 8010010055 8010010065 8010010075 8010010085 8010010095 8010010006 8010010016 8010010026 8010010036 8010010046 8010010056 8010010066 8010010076 8010010086 8010010096 8010010007 8010010017 8010010027 8010010037 8010010047 8010010057 8010010067 8010010077 8010010087 8010010097 8010010008 8010010018 8010010028 8010010038 8010010048 8010010058 8010010068 8010010078 8010010088 8010010098 8010010009 8010010019 8010010029 8010010039 8010010049 8010010059 8010010069 8010010079 8010010089 8010010099 80100100100 80100100101 80100100102 80100100103 80100100104 80100100105 80100100106 80100100107 80100100108 80100100109 801001001000 801001001001 801001001002 801001001003 801001001004 801001001005 801001001006 801001001007 801001001008 801001001009 8010010010000 8010010010001 8010010010002 8010010010003 8010010010004 8010010010005 8010010010006 8010010010007 8010010010008 8010010010009 80100100100000 80100100100001 80100100100002 80100100100003 80100100100004 80100100100005 80100100100006 80100100100007 80100100100008 80100100100009 801001001000000 801001001000001 801001001000002 801001001000003 801001001000004 801001001000005 801001001000006 801001001000007 801001001000008 801001001000009 8010010010000000 8010010010000001 8010010010000002 8010010010000003 8010010010000004 8010010010000005 8010010010000006 8010010010000007 8010010010000008 8010010010000009 80100100100000000 80100100100000001 80100100100000002 80100100100000003 80100100100000004 80100100100000005 80100100100000006 80100100100000007 80100100100000008 80100100100000009 801001001000000000 801001001000000001 801001001000000002 801001001000000003 801001001000000004 801001001000000005 801001001000000006 801001001000000007 801001001000000008 801001001000000009 8010010010000000000 8010010010000000001 8010010010000000002 8010010010000000003 8010010010000000004 8010010010000000005 8010010010000000006 8010010010000000007 8010010010000000008 8010010010000000009 80100100100000000000 80100100100000000001 80100100100000000002 80100100100000000003 80100100100000000004 80100100100000000005 80100100100000000006 80100100100000000007 80100100100000000008 80100100100000000009 801001001000000000000 801001001000000000001 801001001000000000002 801001001000000000003 801001001000000000004 801001001000000000005 801001001000000000006 801001001000000000007 801001001000000000008 801001001000000000009 8010010010000000000000 8010010010000000000001 8010010010000000000002 8010010010000000000003 8010010010000000000004 8010010010000000000005 8010010010000000000006 8010010010000000000007 8010010010000000000008 8010010010000000000009 80100100100000000000000 80100100100000000000001 80100100100000000000002 80100100100000000000003 80100100100000000000004 80100100100000000000005 80100100100000000000006 80100100100000000000007 80100100100000000000008 80100100100000000000009 801001001000000000000000 801001001000000000000001 801001001000000000000002 801001001000000000000003 801001001000000		

**“We’re auditing you.”**



## agency

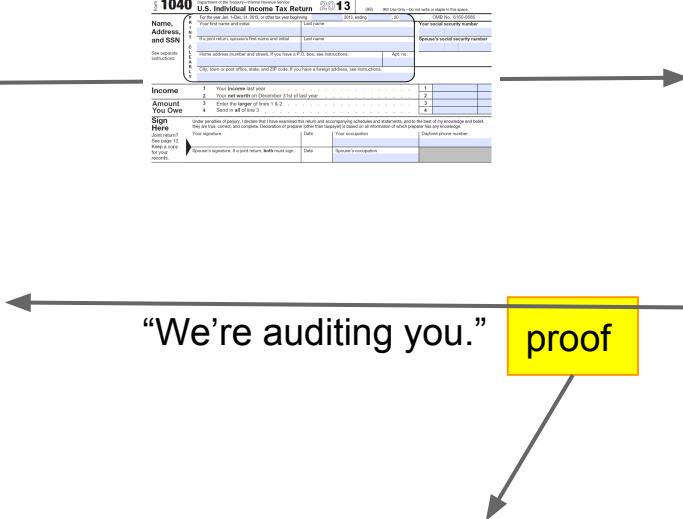
# *Accountable Tax Audit*



subject

agency

# *Accountable Tax Audit*



subject

agency



overseer

# *Accountable Tax Audit*



A photograph of a 2013 U.S. Individual Income Tax Return (1040) form. A large rectangular area in the center of the form is highlighted in red and contains the text "private data".



subject

"We're auditing you."

proof

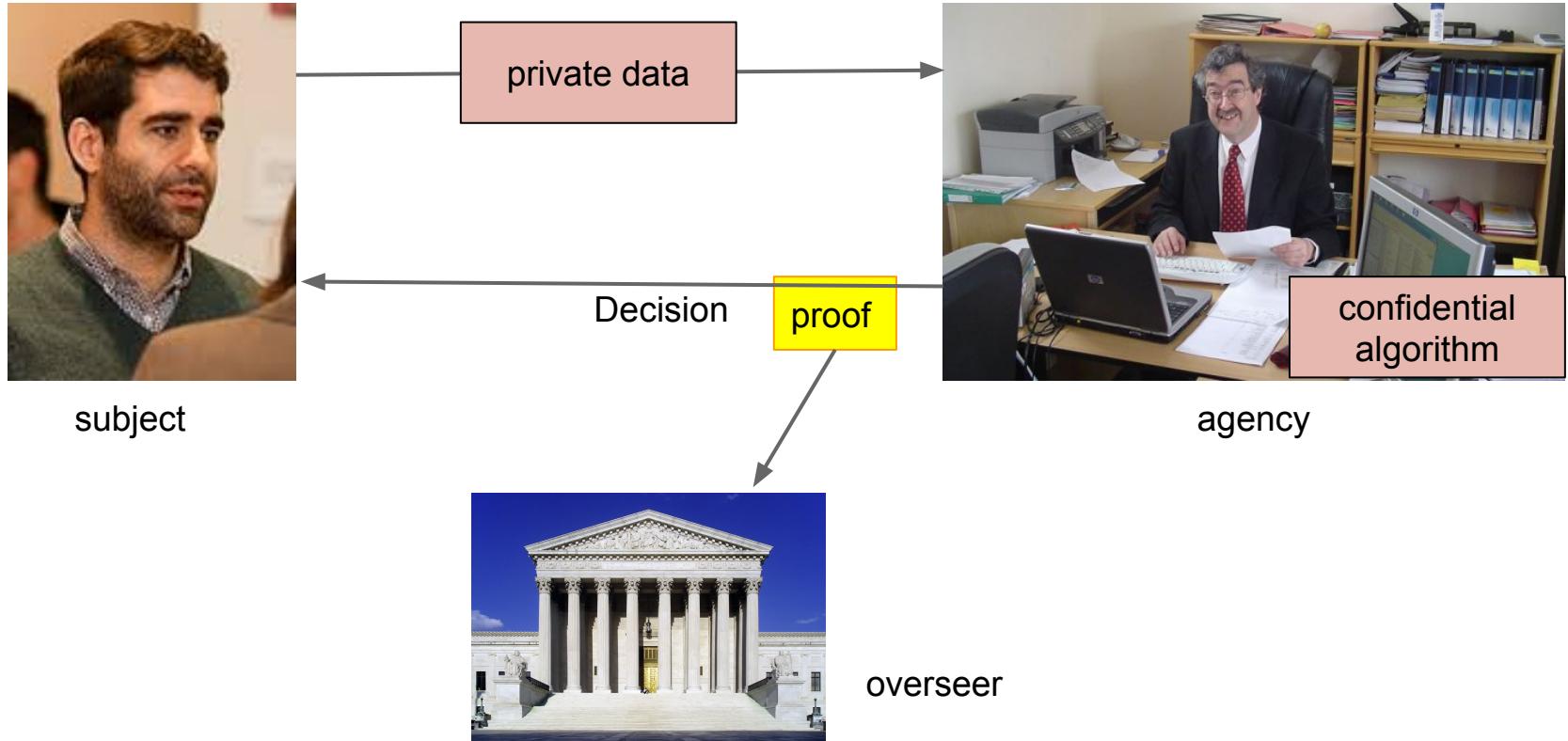
confidential algorithm

agency



overseer

# *Accountable Algorithms*



# Accountable Algorithms: Examples

- Tax audit
- School admission lottery
- Eligibility for housing, credit, ...
- Scoring / predictive decisionmaking
- Search at airport security
- Immigration lottery
- Voting

# Guarantees

Subject verifies that:

- Agency committed to algorithm in advance
- Agency used same algorithm for everyone
- Correct subject-specific data was used
- Randomness was fairly generated

Overseer determines whether:

- Algorithm is lawful, or fair, or good policy

# Role of the Overseer

Overseer: court, executive, or legislature

Can rule on algorithm's suitability, after the fact

Don't have a specification

Model as an oracle, can consult only rarely

# Confidentiality + Accountability

Key idea:

Advanced crypto lets us “square the circle” and keep information confidential while proving it has certain properties.

Goal: Build a working prototype.

# The Protocol: Setting

- Parties: *Agency*, *Subjects*, *Overseer*
- *Agency's policy* is divided into a public *function f* and a secret *parameter y*.
  - Decision for subject  $i$  is  $z_i = f(y, x_i)$  for each subject's data  $x_i$
- *Agency* writes signed protocol output into a *log* (visible to all parties) without interaction by any *subject*.

# The Protocol

- Protocol proceeds in 4 epochs:
  - Commit
  - Randomness
  - Decision
  - Oversight (optional)

# The Protocol



agency

# The Protocol: Normal Operation



agency

signed protocol  
output

subject

Decision



log

# The Protocol: Commit



agency

**ZK-SNARK Public Parameters  
Decision Algorithm  
Commit(Agency input)  
Commit(Privacy randomness)**



log

# The Protocol: Randomness



agency

Random Process (e.g. Beacon, Ceremony, Protocol) output



log

# The Protocol: Decision



agency

Commit(Decision)  
Commit(subject-specific input)  
Proof - ZK-SNARK

subject

Decision

$z$  - Decision  
 $r_2$  - Commitment Randomness



log

# The Protocol: Decision



agency

Commit(Decision)  
Commit(subject-specific input)  
Proof - ZK-SNARK

subject

Decision

$z$  - Decision  
 $r_2$  - Commitment Randomness



log

# The Protocol: Oversight (optional)



overseer



# Major Tools

1. Cryptographic Commitments
2. ZK-SNARKs

# Cryptographic Commitments

A pair of functions

- $C_x \leftarrow \text{Commit}(x, r)$  [r: random]
- $b \leftarrow \text{Verify}(x, r, C_x)$

Such that

- $C_x$  leaks negligible information about  $x$
- $C_x$  is *computationally binding*

# ZK-SNARKs

*Zero-knowledge succinct non-interactive arguments of knowledge*

Given arithmetic circuit with bilinear gates  $\mathbf{C}(\mathbf{x}, \mathbf{a})$ , taking *input*  $\mathbf{x}$  and *witness*  $\mathbf{a}$

Can prove knowledge of  $\mathbf{a}$  such that  $\mathbf{C}(\mathbf{x}, \mathbf{a}) = 1$ . [BCTV14]

Proof is:

non-interactive

zero knowledge w.r.t.  $\mathbf{a}$

succinct:

small constant space (~300 bytes)

small, constant time to check (~ 5 ms)

# ZK-SNARKs

SNARKs exist in a *preprocessing model*

- First, derive from the circuit **C**:
  - a public *proving key*  $\text{pk}$ , and
  - a public *verification key*  $\text{vk}$ .
- From these, anyone can generate and verify proofs.
- Must trust the party who generates  $(\text{pk}, \text{vk})$ .

# How We Use ZK-SNARKs

- $x$ : public information (value)
- $a$ : secret information (witness)
- $C$ : agency circuit
- $C_A(x, a)$  verifies  $x$ ,  $a$ , and result “match up”
  - prover knows how to unwrap commitments, and
  - committed-to values lead to asserted decision
- Zero-knowledge w.r.t. secret information
  - don’t reveal committed-to values
  - can later reveal selectively to appropriate parties

# Extensions

1. Add fairness predicates to  $\mathbf{C}_A$ . Can prove any predicate on agency's secret policy, as part of the zk-SNARK.
2. Less intrusive disputes: structure commitment to subject's private data, so subject can raise dispute without revealing all of subject's private data

# Implementation

- Building a prototype, based on [BCTV14] *libsnark* code for generic ZK-SNARKs.
- Implemented generic  $C_A$ ; experimenting with example applications.

# Example Applications

Classifier

Scoring algorithm

Lottery for immigration, school admission, etc.

# Perspective

Key challenge:

interface between computer science and real policymaking

Computer Science: want complete, precise spec to enforce

Policymaking: compromise-driven,

ex post evaluation of specific policies

# What Computer Science Can Do

enforce procedural fairness

- policy chosen in advance

- same policy for everyone

- fair randomness

enforce “easy parts” of substantive fairness

enable well-informed oversight

# **Fairness vs. Accountability**

Fairness is the ideal.

Accountability is more achievable.

# Accountable Algorithms

Edward W. Felten

Joshua Kroll

Center for Information Technology Policy  
Dept. of Computer Science  
Princeton University

Fairness, Accountability and Transparency in ML Workshop, NIPS 2014  
December 12, 2014

# The Protocol (in handy chart form)

	Agency Writes to Log	Agency Knows Secretly	Agency sends confidentially to each subject
Commit	$\psi$ - ZK-SNARK Parameters f - Decision Algorithm Commit(y - Agency input) Commit( $Q$ - Privacy randomness)	y - <b>Agency input</b> $Q$ - <b>Privacy Randomness</b> $r_1$ - <b>Commitment Randomness</b>	
Randomness	b - <b>Beacon output</b>		
Decision	Commit(z - Decision) Commit(x - subject-specific input) $\Pi$ - ZK-SNARK	$z = f(x, y, b \oplus Q)$ - <b>Decision</b> x - <b>Subject input</b> $r_2$ - <b>Commitment Randomness</b>	z - <b>Decision</b> $r_2$ - <b>Commitment Randomness</b>